NEWSLETTER | The American Philosophical Association

# Philosophy and Computers

SPRING 2018 VOLUME 17 | NUMBER 2







ISSN 2155-9708

APA NEWSLETTER ON

Philosophy and Computers

PETER BOLTUC, EDITOR

VOLUME 17 | NUMBER 2 | SPRING 2018

# FROM THE EDITOR

## *Philosophy in Robotics*

Peter Boltuc

UNIVERSITY OF ILLINOIS SPRINGFIELD AND WARSAW SCHOOL OF ECONOMICS

This note provides an opportunity for reflection on the role of the Committee on Philosophy and Computers as well as this newsletter. It also provides an introduction to this complex, highly interdisciplinary, intergenerational, international and even intercultural issue, which pertains primarily to, broadly defined, philosophy in robotics.

**What is our committee, and the newsletter, all about?** We started in close association with the International Association of Computing and Philosophy (IACAP). The committee was led by Robert Cavalier who starts his 2001 Report from the Chair by saying, "During 2000–2001 the committee sought to investigate and advance the relation between 'philosophy and computers' by working closely with the Steering Committee of the Computing and Philosophy conference in order to encourage the development and expansion of CAP. The PAC committee also sponsored special sessions at the Division Meetings of the APA." The newsletter, led by Jon Dorbolo, published primarily book reviews; it also introduced topics notes in Computer Ethics and a note presenting Herbert A. Simon's work. Some of the tasks were as simple as encouraging some of our colleagues to use email and computers as word processors. But there were already conversations about using automatic proof checkers in teaching critical thinking and logic. There were controversies about the role of online information, but also early stages of conceptual maps, and always abundant problems in computer ethics. I joined this committee in 2003 as a pioneer of e-learning in philosophy. Many of those problems are still present (see the block of five papers on e-learning in philosophy in the fall 2011) though only computer ethics seems to keep its centrality to the field.

**Today, and for the last decade, we seem to be facing slightly bigger challenges, philosophical and social**. The role of AI in our society, as exemplified by the ethics of artificial companions (discussed in past issues of this newsletter by Luciano Floridi 2007, Marcello Guarini 2017), is one of the most tangible philosophical concerns of our times. How should we treat robotic caregivers for children and the elderly, robotic workers, self-driving cars and weapon systems, even robotic lovers? Other philosophical issues include the ontology of virtual beings (Lynn Rudder Baker, 2018; Amie L. Thomasson, 2008; Roxanne M. Kurtz 2009, 2017), ontology of the net (Harry Halpin, 2008; László Ropolyi, 2018) and even computer art (Dominic Mciver Lopes, 2009). We face the need of phenomenology for conscious machines (Gilbert Harman 2007, 2008; Stan Franklin, Bernard Baars, and Umma Ramamurthy, 2007; Igor Aleksander, 2009; P. Boltuc, 2014), computerized epistemology (Jean-Gabriel Ganascia, 2008), or metaphysical foundations for information ethics (Terrell W. Bynum, 2008). Those are the kinds of topics barely ever tackled by strictly philosophical journals, and are rarely present at the APA meetings, outside of the session organized by this committee—since they are essentially interdisciplinary, closely related both to philosophy and also AI.

It may seem that there must be new, vibrant journals in this domain. But in fact, the only journal that covers a similar area is *Minds and Machines*, which started in 1991 and is primarily focused on Artificial Intelligence, and *Ethics and Information Technology*, which started in 1999; there are also a couple of well-established journals in philosophy of engineering. Yet, both the committee and this newsletter are facing certain problems. One of the shortcomings of APA Newsletters that—after the reform of APA website in 2013, which deleted access to single articles—publications in our newsletters are practically nonsearchable by standard web engines. This is a problem, especially since we have some **legacy articles** worth broad attention, such as two original articles by the late **Jaakko Hintikka**: his "Function Logic and the Theory of Computability," published in the fall of 2013, and "Logic as a Theory of Computability," fall 2011, and **John Pollock's** 'Probabilities for AI," published posthumously, thanks to the initiative of Terry Horgan, who was searching for a prestigious open-access publication for this final masterpiece of Pollock's distinguished career.

Those and many other issues standing in front of the committee, and this newsletter, are in need of discussion. I would like to invite members of this committee (past and present), as well as the readers, to engage in this debate, and to send me your contributions to my email epetebolt@gmail.com.

***

The current issue of the newsletter exemplifies many aspects of the breath and the scope of this committee, thus of the newsletter. We open with the article by Jean-Gabriel Ganascia, Catherine Tessier, and Thomas M. Powers (the former chair of this committee) that examines the threat

posed by the so-called killer robots. The article is related to An Open Letter from AI & Robotics Researchers on Killer Robots, promoted by Elon Musk among many others. The authors share some of the concerns by the signatories of that now well-known open letter; they also point out the number of open questions and conceptual issues in need of clarification. The paper is a call for further discussion of this important topic in military ethics.

Then we present the article New Developments in the LIDA Model by Stan Franklin and his team. Several graduate students and researchers wonder about recent progress of this important cognitive architecture that allows AI to exhibit many of the functionalities of human brain. This is a great informal presentation of those developments, appropriate for philosophers, that covers a number of philosophical topics such as motivations, action and language communication. I find the most interesting the section about the self, where LIDA cognitive architecture follows Shaun Gallagher's (2013) pattern theory of the self.

After those two iconic articles, we have two papers by beginning scholars. Jonathan R. Milton follows up on the article by Troy D. Kelley and Vladislav D. Veksler, "Sleep, Boredom, and Distraction—What Are the Computational Benefits for Cognition?" featured in the fall 2015 issue of this newsletter. In his paper, "Distraction and Prioritization: Combining Models to Create Reactive Robots," Milton provides a more applied instrumentation of Kelley and Veksler's idea that "distractability" is sometimes a beneficial feature for a robot; he also singles out some broader philosophical questions. LIDA turns out to be one of the three main cognitive architectures used for the task. In one of the most controversial papers published in this newsletter, Sky Darmos argues that quantum erasers can be used to test animal/robot consciousness. The paper violates a few dogmas of contemporary quantum physics harking back on the state of the theory from *circa* 1950s. At the very least, the paper provides an interesting conceptual possibility how quantum effects, under the traditional Bohr interpretation, could have been used to diagnose consciousness in animals (and, today, in robotic cognitive agents).

We follow up with the paper by Pentti Haikonen, who summarizes the main argument from his recent Finnish-language book devoted to "a new explanation for phenomenal consciousness." Interestingly, Haikonen touches on "the detection problem," but unlike Darmos, the author argues that "the actual phenomenal inner experience cannot be detected as such by physical means from outside; it is strictly personal and subjective." In much of his argument, Haikonen zeroes in on the physical interpretation of qualia. Simon Duan also tackles the issue of unification of consciousness and matter within a metacomputational framework. The author proposes a model that assumes the existence of an operating computer in Platonic realm. The physical universe and all of its contents are modeled as processing output of the Platonic computer.

Next is a paper by László Ropolyi, which uses an Aristotelian framework for building philosophy. The author uses very divergent philosophical traditions that include not only Aristotle but also phenomenology and postmodernism.

Last but not least, Jean-Paul Delahaye and Clément Vidal argue that "that random complexity and organized complexity" are two distinct concepts. By introducing the framework of evolutionary history of the universe the authors attempt to attain a "general measure of complexity." This seems like an important step not only in the theory of complexity, but also in philosophical debate, for instance on Luciano Floridi's non-standard notion of entropy.

***

Different readers may find different articles in this issue interesting, even fascinating, or deeply disturbing, not worth attention. We have iconic AI figures from the US and France; experts (as well as beginning scholars) in computer ethics, theory of computability, or machine consciousness from France, USA, Finland, Belgium, China, Hungary, and the UK. Many top journals struggle with a very low percentage of accepted paper by non-native speakers, ranging below 5 percent—and even those are often from just a few countries with very strong English education, such as Germany, Israel, Italy, and Scandinavia. The benefit of our publication is to facilitate dialogue between disciplines, traditions, and also regional discourses. Of course, we need to reject a number of articles, but in some cases we work with the authors on different versions of their work, even for years—sometimes to no avail. I feel bad about a noted author from India whose paper went for several rewrites but discourse-specificity, and some of the pre-argumentative givens, seemed overly hard to fit with the general discourse of philosophy. There are always challenges and judgment calls to be made. Yet, interdisciplinary and intercultural dialogue allowed on our forum seems rare and hard to replicate. I find it refreshing how computer scientists try to handle centuries-old philosophical problems with different means while we philosophers may sometimes be able to provide a brainstorming kind of feedback for AI experts and programmers.

# FROM THE CHAIR

Marcello Guarini
**UNIVERSITY OF WINDSOR**

## THE 2017 BARWISE PRIZE GOES TO JACK COPELAND

We are pleased to announce that the APA Committee on Philosophy and Computers has awarded the 2017 Barwise Prize to Jack Copeland. Professor Copeland is the world-wide expert on Alan Turing and a leading philosopher of AI, computing and information. He is an author of influential books (2017, 2013, 2012, 2010, 2006, 2005, 2004, 1996, 1993). He has published over a hundred articles, including pioneering work on hypercomputing, which is based on Turing's work but goes far beyond it. He authored the

influential entry, "The Church-Turing Thesis" for the *Stanford Encyclopedia of Philosophy*.

Jack is Distinguished Professor of Philosophy, and Department Head, at the University of Canterbury, New Zealand, where he is Director of the Turing Archive for the History of Computing. He is co-founder and Co-Director of the Turing Centre Zürich (TCZ) at the Swiss Federal Institute of Technology (ETH Zürich), where he is a permanent International Fellow. He is also Honorary Research Professor at the University of Queensland in Australia. He has been a visiting professor at a number of top universities world-wide and keynote speaker at numerous major conferences in the areas of Philosophy and Computing and Philosophy and Cognitive Science. In 2016, he received the international Covey Award, recognizing "a substantial record of innovative research in the field of computing and philosophy."

In terms of his direct connections to the APA Philosophy and Computers Committee, Jack co-organized with this committee the 2005 and 2006 meeting of the Society for Machines and Mentality at the APA. At the 2005 session he gave a paper entitled "Ontic versus epistemically embedded computation."

### CURRENT ACTIVITIES OF THE COMMITTEE

As well as deliberating over the Barwise Prize, the Philosophy and Computers Committee has been busy organizing sessions for the 2018 Central and Pacific APA meetings. As was announced in the previous edition of our newsletter, committee member Peter Boltuc chaired a session at the Central APA in February, and Fritz McDonald will be chairing a session at the Pacific APA in March.

Readers of the newsletter are encouraged to contact the committee chair (Marcello Guarini, mguarini@uwindsor.ca) if they are interested in proposing a symposium at the APA that engages any of the wide range of issues associated with philosophy and computing. We are happy to continue facilitating the presentation of high quality research in this area.

As most who are reading this newsletter already know, the weather at the 2018 Eastern APA meeting was not exactly accommodating. Thanks to those who were able to make it to our Barwise Prize session to see the 2016 winner of the award, Ed Zalta, give his talk. Many thanks to everyone involved in making that session happen.

### FUTURE OF THE COMMITTEE

Piotr Boltuc has been elected the next associate chair of the philosophy and computers committee. Piotr's term will begin on July 1, 2018. On July 1, 2019, Piotr will become chair of the committee. Daniel Susser and Jack Copeland will join the committee on July 1, 2018, for two-year terms. Thanks to all three for taking on these responsibilities. Fritz McDonald and Gualtiero Piccinini will be coming to the end of their terms in 2018—many thanks to both of them for all their efforts.

As most of you have heard, the APA board of officers has voted to dissolve the "philosophy and X committees." This includes the philosophy and law committee, the philosophy and medicine committee, and, yes, even our own philosophy and computers committee. The announcement can be found at http://www.apaonline.org/news/388037/Changes-to-APA-Committees.htm.

Our own Piotr Boltuc, in his opening contribution to this issue of the newsletter, makes a very strong case for the continued relevance of the committee. I look forward to continuing to work with Piotr and others to ensure that the issues engaged by our committee continue to be represented in the discourse of the APA. Obviously, many of us hope this takes the form of the APA allowing our committee to exist beyond June 30, 2020—the scheduled phase-out date. Failing that, we hope the interests and concerns of the committee will be included in other committees or APA activities. Please keep looking for our sessions at APA meetings; we have plans to continue organizing them at least through 2020.

## ARTICLES

## *On the Autonomy and Threat of "Killer Robots"*

Jean-Gabriel Ganascia
SORBONNE UNIVERSITY; MEMBER OF THE INSTITUT UNIVERSITAIRE DE FRANCE, CHAIRMAN OF THE CNRS ETHICAL COMMITTEE

Catherine Tessier
ONERA AEROSPACE LAB, FRANCE; INFORMATION PROCESSING AND SYSTEMS DEPARTMENT

Thomas M. Powers
UNIVERSITY OF DELAWARE; DEPARTMENT OF PHILOSOPHY AND CENTER FOR SCIENCE, ETHICS & PUBLIC POLICY

### INTRODUCTION

In the past, renowned scientists such as Albert Einstein and Bertrand Russell publicly engaged, with courage and determination, the existential threat of nuclear weapons. In more recent times, scientists, industrialists, and business leaders have called on states to institute a ban on what are—in the popular imagination—"killer robots." In technical terms, they are objecting to LAWS (Lethal Autonomous Weapons Systems), and their posture seems similar to their earlier, courageous counterparts. During the 2015 International Joint Conference on Artificial Intelligence (IJCAI)—which is the premier international conference of artificial intelligence—some researchers in the field of AI announced an open letter warning of a new AI arms race and proposing a ban on offensive lethal autonomous systems. To date, this letter has been signed by more than 3,700 researchers and by more than 20,000 others, including (of note) Elon Musk, Noam Chomsky, Steve Wozniak, and Stephen Hawking.

In the summer of 2017, at the most recent IJCAI held in Melbourne, Australia, another open letter was presented,

signed by the heads of many companies in the fields of robotics and information technologies, among whom Elon Musk was very active. This second letter urged the United Nations to resume its work toward a ban on autonomous weapons, which had been suspended for budgetary reasons.

It is no doubt incumbent on every enlightened person, and in particular on every scientist, to do everything possible to ensure that the industrialized states give up the idea of embarking on yet another mad arms race, the outcome of which might escape human control. This seems obvious, especially since, according to the authors of these two open letters, we would be at the dawn of a third revolution in the art of war, after gunpowder and the atomic bomb.

If these positions appear praiseworthy at first, should we not also wonder about the actual threats of these lethal autonomous weapon systems? To remain generous and sensitive to great humanitarian causes, should we not also remain rational and maintain our critical sensibilities? Indeed, even though considerable ethical problems arise in the evolution of armaments—from landmines to drones, and recently to the massive exploitation of digitized information and electronic warfare—it appears on reflection that this third revolution in the art of war is very obscure. Where the first two revolutions delivered considerable increases in firepower, we find here an evolution of a very different order.

Moreover, the so-called "killer robots" that have been the targets of three years of numerous press articles, open letters, and debates seem to be condemned by sensational and anxiety-laced arguments, mostly to the exclusion of scientific and technical ones. The term "killer robot" suggests a robot that would be driven by the *intention* of killing and would even be *conscious* of that intention, which at this stage in the science does not make sense to attribute to a machine—even one that has been designed for destroying, neutralizing, or killing. For instance, one does not speak of a "killer missile" when it happens that a missile kills someone. "Killer robot" is a term that is deployed for rhetorical effect, that works to hinder ethical discussion, and that aims at manipulating the general public. Do the conclusions of these arguments also hold against "killing robots"? Is there an unavoidable technological path from designing "killing robots" to deploying "killer robots"?

To get a better understanding of these questions, we aim here to put forward a detailed analysis of the 2015 open letter, which was one of the first public manifestations of the desire to ban LAWS. Our reservations concerning the declarations that this letter contains should help to open the scientific and philosophical debates on the controversial issues that lie at the heart of the matter.

## THE ARGUMENT FOR A BAN

The 2015 open letter was revealed to journalists and, by extension, to a broad audience during the prestigious IJCAI in Buenos-Aires, Argentina. In its first sentence, the letter warned that "[a]utonomous weapons select and engage targets without human intervention," and concluded after four short paragraphs by calling for a ban on offensive forms of such weapons. This public announcement had been preceded by an invitation for signatories within the AI scientific community and beyond, including a wider community of researchers, technologists, and business leaders. Many of the most prominent AI and robotics researchers signed it, and outside the AI community many prominent people brought their support to this text. Initially, the renown and humanitarian spirit of the co-signers may have inclined many people to subscribe to their cause. Indeed, the possibility of autonomous weapons that select their targets and engage lethal actions without human intervention appears really terrifying.

However, after a careful reading of the first open letter, and in consideration of the subsequent public statements on the same topics—e.g., the IJCAI 2017 (second) open letter and video[1] that circulated widely on the web towards the end of 2017—we think a closer analysis of the deployed arguments clearly shows that the letter raises many more questions than it solves. Despite the fame and the scientific renown of the signatories, many statements in the letter seem to be questionable from a scientific point of view. In addition, the text encompasses declarations that are highly disputable and that will certainly be belied, very soon, by upcoming technological developments. These are the reasons why, as scientists and experts in the field, it seems incumbent upon us to scrutinize the claims that these public announcements contain and to re-open the debate. We are not disparaging the humanitarian aims of the authors of the letter; we do, however, want to look more closely at the science and the ethics of this issue. Even though we share the same feeling of unease that has likely motivated the authors and the signatories of these open letters, we want to bring into focus where, we believe, the scientific case is lacking for the normative conclusion they draw.

For ease of reference, the content of the 2015 Open Letter has been appended to this article, with numbered lines added to facilitate comparison between our text and theirs.

The first paragraph (l. 10–17) describes recent advances in artificial intelligence that will usher in a new generation of weapons that qualify as autonomous because they "*select and engage targets without human intervention.*" These weapons will possibly be deployed "*within years, not decades*" and will constitute "*the third revolution in warfare, after gunpowder and nuclear arms.*" The next paragraph (l. 18–33) explains why a military artificial intelligence arms race would not be beneficial for humanity. The two main arguments are, first, that "*if any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable,*" and second, as a consequence, "*autonomous weapons will become the Kalashnikovs of tomorrow*" (i.e., they will become ubiquitous because they will be cheap to produce and distribution will flow easily from states to non-state actors). In addition, this paragraph warns that autonomous weapons are "*ideal*" for dirty wars (i.e., "*assassinations, destabilizing nations, subduing populations and selectively killing a particular ethnic group*"). The third paragraph (l. 34–40) draws a parallel between autonomous weapons and biological or chemical weapons, the development of which most scientists have rightly shunned. AI researchers, it is implied, would "*tarnish*

*their field"* by developing AI weapons. Finally, the last paragraph (l. 41–44) summarizes the content of the letter and then calls for a ban on offensive autonomous weapons.

Our perplexity comes from these four aspects of the general argument, as developed in the letter:

1) The notion of "autonomous weapon" that motivates the letter is obscure; its novelty and what distinguishes it from AI weapons in general are sources of confusion. At least this much is certain: not all AI weapons are autonomous, according to the definition given by the authors (selecting and engaging targets without human intervention). Contrary to what is claimed, the technical feasibility of autonomous weapons deployment in the near future is far from obvious.

2) Despite the dramatic illustrations given in the letter and repeated in the video to which we referred above, the specific noxiousness of autonomous weapons that makes them "*ideal*" for dirty military actions and that differentiates them from current weapons is not obvious from a technical point of view.

3) The analogy between the current attitude of AI scientists faced with the development of autonomous weapons and the past attitude of scientists faced with the development of chemical and biological weapons is far from clear. Besides, the parallel between the supposed outbreak of autonomous weapons in contemporary military theaters and the advent of gunpowder or nuclear bombs in warfare is highly debatable.

4) Lastly, the ban on offensive autonomous weapons is not new and is already being discussed by military leaders themselves, which makes this declaration somewhat irrelevant.

The remainder of this article is dedicated to a deeper analysis of the four points above.

## AUTONOMOUS WEAPONS

What exactly is the notion of "autonomous weapon" to which the letter refers? Autonomy is the capability for a machine to function independently of another agent (human, other machine) exhibiting non-trivial behaviors in complex, dynamic, unpredictable environments.[2] The autonomy of a weapon system would involve sensors to assist in automated decisions and machine actions that are calculated without human intervention. Understood in this way, autonomous weapons have already existed for some time, as exemplified by a laser-guided missile that "hangs" a target.

The current drones that are operated and controlled manually at more than 3,000 km from their objectives use such autonomous missiles. If this were the meaning of "autonomous weapons" in this letter, the notion would correspond only to a continuous progression in military techniques. In other words, this would just be an augmentation in the distance between the "soldier" (or, more precisely, the operator) and its target. In this respect, among a bow and arrow, a musket, a gun, a canon, a bomber, and a drone, there is just a difference in the order of magnitude of the arms' ranges. However, the text of the open letter does not say this, but rather claims that (l. 10) *[a]utonomous weapons select and engage targets without human intervention*. The question, then, is not about the range of action but about the "logical" nature of the weapon: until now, and for centuries, a human soldier aimed at the target before firing, while in the future, with autonomous weapons, the target will be abstractly specified in advance. In other words, the mode of designating the target changes. While up to now, the objective, i.e., the target, was primarily an index on which the human aimed, in the near future it will just become an abstract symbol designated by a predefined rule. Since no human is involved in triggering the lethal action, this evolution of warfare seems terrifying, which would justify the concerns of the open letter.

Let us note that the concept of "autonomy" is problematic, firstly because various stakeholders (among them, scientists) give the term multiple meanings.[3] An "autonomous weapon" can thus designate a machine that reacts automatically to certain predefined signals, that optimizes its trajectory to neutralize a target for which it has automatically recognized a predefined signature, or that automatically searches for a predefined target in a given area. Rather than speaking of "autonomous weapons," it seems more relevant to study which functions are or could be automated, which is to say, delegated to computer programs. Further, we should want to understand the limitations of this delegation, in the context of a sharing of authority (or control) with a human operator, which sharing may vary during the mission.

Guidance and navigation functions have been automated for a long time (e.g., automatic piloting) and have not raised significant questions. These are non-critical operational functions. But automatic identification and targeting are more sensitive functions. Existing weapons have target recognition capabilities based on predefined models (or signatures): the recognition software matches the signals received by the sensors (radar signals, images, etc.) with its signature database. This recognition generally concerns large objects that are "easy" to recognize (radars, airbases, tanks, missile batteries). But the software is unable to assess the situation around these objects—for example, the presence of civilians. Targeting is carried out under human supervision, before and/or during the course of the mission.

## *INELUCTABILITY*

The authors seem to suggest that this evolution is ineluctable because, if specification of abstract criteria and construction of the implementing technology is cheaper and faster than recruiting and training soldiers, and assuming that modern armies have the financial and technical wherewithal to make these weapons, then autonomous weapons will eventually predominate. This complicated point deserves some more in-depth analysis, since the definition of the *criteria* to which the open letter refers appears sometimes

very problematic, despite the progress of AI and machine learning techniques. Many problems remain to be solved. For instance, how will the technology differentiate enemies from friends in asymmetric wars, where the soldiers don't wear uniforms? More generally, when humans are not able, on the basis of a given set of information, to discriminate cases that meet criteria from cases that don't, how will machines do better? If humans cannot discern, from photos, which are the child soldiers and which are children playing war, it is illusory to hope to build a machine that automatically learns these criteria, on the basis of the same set of information. Will algorithms be able to recognize a particular individual from their facial features, a foe from their military uniform, a person carrying a gun, a member of a particular group, a citizen of a particular country whose passport will be read from a remote device? It will be impossible to build a training set.

In recognition of these remaining problems, it seems that the supposed ineluctability of the evolution that would spring from the AI state of the art is debatable, and certainly not *"feasible within a few years"* as the letter claims. It would have been more helpful had the authors of the letter elaborated on what precisely will be feasible in the near future, especially as far as automated situation assessment is concerned. The assertion that full-blown autonomous weapons are right around the corner would then have been placed in context.

### *ON THE FORMAL SPECIFICATIONS OF AUTONOMY*

Current discussions and controversies focus on the fact that an autonomous weapon would have the ability to recognize complex targets in situations and environments that are themselves complex and would be able to engage (better than can humans) such targets on the basis of this recognition. Such capabilities would suppose the weapon system has the following abilities:

- to have a formal (i.e., mathematical) description of the possible states of the environment, of the elements of interest in this environment, and of the actions to be performed, even though there is no "standard situation" or environment

- to recognize a given state or a given element of interest from sensor data

- to assess whether the actions that are computed respect the principles of humanity (avoid unnecessary harms), discrimination (distinguish military objectives from populations and civilian goods), and proportionality (adequacy between the means implemented and the intended effect) of the International Humanitarian Law (IHL)

Issues of a philosophical and technical nature are related to the ability of the system to automatically "understand" a situation, and in particular to automatically "understand" the intentions of potential targets. Today, weapon system actions are undertaken with human supervision, following a process of assessment of the situation, which seems difficult to formulate mathematically. Indeed, the very notion of agency, when humans and non-human systems act in concert, is quite complicated and also fraught with legal peril.

Beyond the philosophical and technical aspects, another issue is whether it is ethically acceptable that the decision to kill a human being, who is identified as a target by a machine, can be delegated to this machine. More specifically, with respect to the algorithms of the machine, one must wonder how and by whom the characterization, model, and identification of the objects of interest would be set, as well as the selection of *some* pieces of information (to the exclusion of some others) to compute the decision. Moreover, one must wonder who would specify these algorithms and how it would be proven that they comply with international conventions and rules of engagement. And as we indicated above, the accountability issue is central: Who should be prosecuted in case of violation of conventions or misuse? It is our contention that these difficult formal issues will delay (perhaps indefinitely) the advent of the sort of autonomous weapons that the authors so fear.

Finally, it is worth noting that the definition of autonomous weapons (*Autonomous weapons select and engage targets without human intervention* (l. 10)) comes from the 2012 US Department of Defense Directive Number 3000.09 (November 21, 2012. Subject: Autonomy in Weapon Systems). Nevertheless, the authors of the letter have truncated it. As a matter of fact, the complete definition given by the DoD directive is the following: *Autonomous weapon system: a weapon system that, once activated, can select and engage targets without further intervention by a human operator. This includes human-supervised autonomous weapon systems that are designed to allow human operators to override operation of the weapon system, but can select and engage targets without further human input after activation*.

From the DoD directive, one learns in particular that (3) "*Autonomous weapon systems may be used to apply non-lethal, non-kinetic force, such as some forms of electronic attack, against materiel targets*" in accordance with DoD Directive 3000.3. Therefore, we should bear in mind that a weapon (in general) should be distinguished from a lethal weapon. Indeed, a weapon system is not necessarily a system that includes lethal devices.

Hence, the proffered, alarming example of what autonomous weapons technology could bring—"*armed quadcopters that can search for and eliminate people meeting certain pre-defined criteria*" (l. 11–12)—seems more fitting for the tabloid press. For this example to be taken seriously, some of those targeting criteria should be made explicit, and current and future technology should be examined as to whether a machine would be able to assign instances to criteria, with no uncertainty, or with less uncertainty than a human assessment. For example, the criterion "target is moving"—for which no AI or autonomy is required—is very different from the criterion "target looks like this sketch and attempts to hide."

## HARMFULNESS

The second paragraph (l. 18–33) is mainly focused on the condemnation of automated weapons.

### *THE ETHICS OF ROBOT SOLDIERS*

From the beginning, this paragraph seems intended to measure the costs and benefits of autonomous weapons, but it proceeds too quickly by dismissing debates about the possible augmentation or diminution of casualties with AI-based weapons. While the arguments for augmentation rely upon the possible multiplication of armed conflicts, the arguments for diminution seem to be based on the position of the roboticist Ronald Arkin.[4] According to Arkin, robot soldiers would be more ethical than human soldiers because autonomous machines would be able to keep their "blood cold" in any circumstance and to obey the laws of the conduct of a just war. Note that this argument is suspect because the relevant part of just war laws—the conditions for just conduct or *jus in bellum*—are based on two further principles. As we indicated above, the principle of *discrimination*, according to which soldiers have to be distinguished from civilians, and the principle of *proportionality*, which limits a response to be proportional to the attack, are both crucial to building an ethical robot soldier. Neither discrimination nor proportionality can be easily formalized, so it is unclear how robot soldiers could obey the laws of just war. The problem is that, as mentioned in the previous section, there is no obvious way to extract concrete objective criteria from these two abstract concepts. However, interestingly, the open letter never mentions this formal problem, even though it could help to reinforce its position against autonomous weapons.

### *IDEAL WEAPONS FOR DIRTY TASKS*

The main argument concerning the harmfulness of autonomous weapons is that they "*are ideal for tasks such as assassinations, destabilizing nations, subduing populations, and selectively killing a particular ethnic group.*" The different harms belonging to this catalog appear to be highly heterogeneous. What is common to these different goals? Further, the adjective "*ideal*" is particularly obscure. Does it mean that these weapons are perfectly appropriate for the achievement of those dirty tasks? If that is the case, it would have helped to give more details and to show how autonomous weapons would facilitate the work of assailants. Such an elaboration would have been important because, at first glance, there is no evidence that autonomous weapons will be more precise than classical weapons (e.g., drones) for assassination or selective killing of a particular ethnic group. Indeed, it is difficult to imagine how autonomous machines could select, more efficiently than other weapons, the individuals that are to be killed, or discern expeditiously members of human groups, depending on their race, origin, or religion. Finally, the underlying premise of the "harmfulness" argument is worth questioning, for it is not clear that those conducting "dirty wars" care much about precision or selectivity. Indeed, this "not caring" may be a central trait of the "dirtiness" of such aggression.

### *NECESSARY DISTINCTIONS*

Underlying the discussion of these loosely related "dirty" tasks and a possible arms race, there is a confusion between three putative properties of autonomous weapons that, taken one by one, are worth discussing: firepower, precision, and diffusion. Despite the reference to gunpowder and nuclear weapons (l. 16–17, 24, 40), there is no direct relation between autonomy of arms and their firepower. Further, it is not any more certain that autonomous weapons would reach their targets more precisely than classical weapons. The series of "drone papers"[5] shows how difficult it is to systematize human targets selection and to automatically gather exact information on individuals by screening big data. Lastly, the argument about the diffusion of autonomous weapons is in contradiction with the supposed specific role of major military powers in autonomous weapon development. More precisely, the problem appears when we consider the following claims:

1) *If any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable* (l. 21–23), (which we consider to be probable)

2) *autonomous weapons will become the Kalashnikovs of tomorrow* (l. 24), (which is also possible)

However, even if claims 1 and 2 above are plausible separately, they seem jointly implausible. (By comparison, the development of nuclear weapons *did* start an arms race, but also kept nuclear armaments out of the hands of all but the "nuclear club" of nations.) There may even be an antinomy between 1 and 2, because if *only* major military powers would be able to promote scientific programs to develop autonomous weapons, then it is likely that these scientific programs would be too costly to develop for industries, without rich state support, or for poor countries or non-state actors, which means that these arms couldn't so quickly become sufficiently cheap that they would spread throughout all humankind. Some weapons might be more easily replicated, once information technologies have been developed, and military powers could act as pioneers in that respect. However, nowadays, it appears that military industries are not guiding technical development in information technologies, as was the case in the twentieth century (at least until the end of the seventies), but that more often the opposite is the case: information technology industries (and dual-purpose technologies) are ahead of the military technologies. Undoubtedly, information technology industries would become prominent in developing autonomous weapons technologies if there were a mass market for autonomous weapons, as the authors of this open letter assume. Lastly, if these technologies were potentially so cheap that they could be spread widely, there would be a strong incentive for the major military powers to keep "a step ahead" to ensure the security of their respective populations.

The paragraph ends with a rather strange sentence (l. 32–33): "*There are many ways in which AI can make battlefields safer for humans, especially civilians, without creating new tools for killing people*." This suggests that AI would benefit defense whereas autonomous weapons would not. Nevertheless, what has been argued previously against autonomous weapons can fit all other AI applications in

defense in the same way. Moreover, and to add to the confusion in this claim, the terms *autonomous weapon* (l. 10, 15, 18, 24, 29, 43), *AI weapon* (l. 22, 35) and *AI arms* (l. 21, 31, 42) seem, for the authors, to be interchangeable or synonymous phrases. Yet equipping a weapon, whether lethal or not, with some AI (e.g., a path-planning function) does not necessarily make it autonomous, and conversely, some forms of autonomy (e.g., an autopilot) may hinge on automation without involving any AI.

## ANALOGIES WITH OTHER WEAPONS

A third central claim in the general argument concerns military analogies with other weapons: nuclear weapons on the one hand and biological and chemical weapons on the other. All of these parallels are troublesome.

### *THIRD REVOLUTION IN WARFARE*

It is announced (l. 15–17) that the development of autonomous weapons would correspond to a third revolution in warfare, after gunpowder and nuclear weapons. Later, the analogy with nuclear weapons is repeated twice (l. 24 and l. 40) in order either to draw connections or to underline differences. Based on our observations above, it does not seem that autonomous weapons will lead to an augmentation in firepower but, instead, to an increase in the distance between the soldier and his/her target. If there is something innovative in autonomous weaponry, it is in *range* rather than *power*. Therefore, it would have been better to compare autonomous weapons with the bow and arrow, the musket, or the bomber drone instead of with weapons for which incidence range is totally heterogeneous.

### *PARALLEL WITH CHEMICAL AND BIOLOGICAL WEAPONS*

The third paragraph draws a parallel between autonomous weapons and weapons that have been considered morally repugnant, such as the chemical and biological weapons that scientists don't develop anymore, because they "have no interest in building" them, and they "do not want others to tarnish their field by doing so" (l. 34–36).

The comparison is questionable. Indeed, historically, it is mostly German and French chemists who developed many chemical weapons (mustard gas, phosgene, etc.) during the Great War. Similarly, Zyklon B had been conceived by Walter de Heerdt, a student of Fritz Haber, recipient of Nobel Prize in Chemistry, as a pesticide. The ban on chemical and biological weapons did not spring from scientists but from the collective consciousness, after the First World War, of the horrors of their use.

In a somehow different register, the scientific community didn't oppose, as a whole, the development and deployment of nuclear weapons. The presence of a large number of great physicists in military nuclear research centers attests to this fact.

In terms of the parallel, it is far from clear that AI will lead to autonomous weapons, and far from clear that autonomous weapons will be widely viewed as morally abhorrent, compared to the alternatives.

## THE BAN CLAIM

### *A BAN ON OFFENSIVE AUTONOMOUS WEAPONS*

The final paragraph proposes a "*ban on offensive autonomous weapons beyond meaningful human control*" (l. 43–44). Nonetheless, the authors should know that many discussions have already taken place, that scientists have barely participated in these discussions, and that in the United States, in 2012, the Defense Department already decided on a moratorium on the development and the use of autonomous and semi-autonomous weapons for ten years (see above reference to the DoD Directive 3000.09). For several years, the United Nations has also been concerned about this issue. It is therefore difficult to understand the exact position of the scientific authors of the letter, especially if it does not invoke the debates that have already taken place, and to the extent that it relies on some not-altogether-germane considerations—precision, ubiquity, illicit use, firepower, etc.—such as we have explained above.

In short, the conclusion of a ban does not seem to be justified by the general argument of the letter (given the problems we have noted), nor by the novelty of the position they are staking out. There is a ban, and states are not racing ahead to deploy offensive, lethal, autonomous weapons systems. But might we be missing something? Might the authors foresee a deeper reason for scientists and technologists to eliminate the *very possibility* of an unlikely but terrifying threat?

Such would be the conclusion of an argument from the "precautionary principle," which could be the motivating principle of the ban. The precautionary principle is often invoked in environmental ethics, especially in assessing geo-engineering to combat climate change. The idea is that, while new technologies promise benefits, the threat of them going astray is so cataclysmic in terms of their costs that we must act to eliminate the threat, even when the likelihood of cataclysm is very small. The imagined threat here would be the continued development of autonomous weapon systems leading to a military AI arms race, or the mass proliferation of AI weapons in the hands of unscrupulous non-state actors, as the authors of the open letter envision.

Wallach and Allen discussed a similar argument against AI in their 2009 book *Moral Machines*:[6]

> The idea that humans should err on the side of caution is not particularly helpful in addressing speculative futuristic dangers. This idea is often formulated as the "precautionary principle" that if the consequences of an action are unknown but are judged to have some potential for major or irreversible negative consequences, then it is better to avoid that action. The difficulty with the precautionary principle lies in establishing criteria for when it should be invoked. Few people would want to sacrifice the advances in computer technology of the past fifty years because of 1950s fears of a robot takeover.

In answer to the "precautionary" challenge to autonomous weapons, it seems that Wallach and Allen provide the

right balance between ethical concern and scientific responsibility:

> The social issues we have raised highlight concerns that will arise in the development of AI, but it would be hard to argue that any of these concerns leads to the conclusion that humans should stop building AI systems that make decisions or display autonomy. [. . .] We see no grounds for arresting research solely on the basis of the issues presently being raised by social critics or futurists.

### *SCIENTIFIC AUTHORS*

Let us end by going to the beginning—with a consideration of the title (l. 8–9): *"Autonomous Weapons: An Open Letter from AI & Robotics Researchers."*

Who exactly are the AI and Robotics Researchers who wrote the open letter? As a matter of fact, nothing in their presentation allows those who wrote the letter to be distinguished from those who have signed it. The question is all the more important, as some tensions within the arguments of the text suggest that some negotiations took place. In any case the open letter cannot appear as coming from *all* AI and robotics researchers. Some members of this community, both in Europe and in the United States—not to mention the authors of this present article—have already disagreed with the content of the open letter.

To conclude, scientists and members of the artificial intelligence community may not wish to adhere to the position expressed in the open letter, not because they are interested in developing autonomous weapons or are not "sufficiently humanitarian," but because the arguments conveyed in the letter are not sufficiently grounded in science. We think it is our duty to publicly express our disagreement because when scientists communicate in the public sphere, not as individuals, but as a scientific community as a whole, they must be sure that the state of the art of their scientific knowledge fully warrants their message. Otherwise, such public pronouncements are nothing more than expressions of one opinion among others, and may lead to more misinformation than comprehension—they may generate "more heat than light."

It is also worth sounding another cautionary note here. When scientists decide to take the floor in the public arena, they ought to ensure that their scientific knowledge fully justifies their declarations. In these times, which some commentators have declared as a "post-truth era," the rigor of scientists' arguments is more important than ever in order to fight fake-news. This can only be ascertained after they engage in debate in their respective scientific communities, especially when some of their colleagues are not in agreement with them. Otherwise, without such open dialogue—discussions which are crucial in scientific communities to establish claims of knowledge—the public may come to doubt future declarations of scientists on ethical matters, especially if they concern technological threats. Any scientific pronouncement, whether meant for an expert community or addressed to the public, ought to take utmost care to preserve scientific credibility.

## APPENDIX

Embargoed until 4PM EDT July 27 2015/5PM Buenos Aires/6AM July 28 Sydney
This open letter will be officially announced at the opening of the IJCAI 2015 conference on July 28, and we ask journalists not to write about it before then. Journalists who wish to see the press release in advance of the embargo lifting may contact Toby Walsh.
Hosting, signature verification and list management are supported by FLI; for administrative questions about this letter, please contact tegmark@mit.edu.

### Autonomous Weapons: An Open Letter from AI & Robotics Researchers[7]

Autonomous weapons select and engage targets without human intervention. They might include, for example, armed quadcopters that can search for and eliminate people meeting certain pre-defined criteria, but do not include cruise missiles or remotely piloted drones for which humans make all targeting decisions. Artificial Intelligence (AI) technology has reached a point where the deployment of such systems is—practically if not legally—feasible within years, not decades, and the stakes are high: autonomous weapons have been described as the third revolution in warfare, after gunpowder and nuclear arms.

Many arguments have been made for and against autonomous weapons, for example that replacing human soldiers by machines is good by reducing casualties for the owner but bad by thereby lowering the threshold for going to battle. The key question for humanity today is whether to start a global AI arms race or to prevent it from starting. If any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable, and the endpoint of this technological trajectory is obvious: autonomous weapons will become the Kalashnikovs of tomorrow. Unlike nuclear weapons, they require no costly or hard-to-obtain raw materials, so they will become ubiquitous and cheap for all significant military powers to mass-produce. It will only be a matter of time until they appear on the black market and in the hands of terrorists, dictators wishing to better control their populace, warlords wishing to perpetrate ethnic cleansing, etc. Autonomous weapons are ideal for tasks such as assassinations, destabilizing nations, subduing populations and selectively killing a particular ethnic group. We therefore believe that a military AI arms race would not be beneficial for humanity. There are many ways in which AI can make battlefields safer for humans, especially civilians, without creating new tools for killing people.

Just as most chemists and biologists have no interest in building chemical or biological weapons, most AI researchers have no interest in building AI weapons—and do not want others to tarnish their field by doing so, potentially creating a major public backlash against AI that curtails its future societal benefits. Indeed, chemists and biologists have broadly supported international agreements that have successfully prohibited chemical and biological weapons, just as most physicists supported the treaties banning space-based nuclear weapons and blinding laser weapons.

In summary, we believe that AI has great potential to benefit humanity in many ways, and that the goal of the field should be to do so. Starting a military AI arms race is a bad idea, and should be prevented by a ban on offensive autonomous weapons beyond meaningful human control.

## NOTES

1. https://www.youtube.com/watch?v=9CO6M2HsoIA
2. Alexeï Grinbaum, Raja Chatila, Laurence Devillers, Jean-Gabriel Ganascia, Catherine Tessier, and Max Dauchet, "Ethics in Robotics Research: CERNA Recommendations," *IEEE Robotics and Automation Magazine* (January 2017). doi: 10.1109/MRA.2016.2611586.
3. Vincent Boulanin and Maaike Verbruggen, "Mapping the Development of Autonomy in Weapon Systems," Stockholm International Peace Research Institute (SIPRI) (November 2017) https://www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_0.pdf.

   The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems, Version 2. IEEE, 2017, http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.
4. Ronald Arkin, *Governing Lethal Behavior in Autonomous Robots* (Chapman & Hall/CRC Press, 2009).
5. A series of papers published by an online publication ("The Intercept") details the drone assassination program of US forces in Afghanistan, Yemen, and Somalia. Available at https://theintercept.com/drone-papers/.
6. Wendell Wallach and Collin Allen, *Moral Machines: Teaching Robots Right from Wrong* (Oxford University Press, 2009), 52–53.
7. https://futureoflife.org/open-letter-autonomous-weapons/

# *New Developments in the LIDA Model*

Stan Franklin
UNIVERSITY OF MEMPHIS

Steve Strain
UNIVERSITY OF MEMPHIS

Sean Kugele
UNIVERSITY OF MEMPHIS

Tamas Madl
AUSTRIAN RESEARCH INSTITUTE FOR ARTIFICIAL INTELLIGENCE, VIENNA, AUSTRIA

Nisrine Ait Khayi
UNIVERSITY OF MEMPHIS

Kevin Ryan
UNIVERSITY OF MEMPHIS

## INTRODUCTION

Systems-level cognitive models are intended to model minds, which we take here to be control structures[1] for autonomous agents.[2] The LIDA (Learning Intelligent Decision[3] Agent) systems-level cognitive model is intended to model human minds, some animal minds, and some artificial minds, be they software agents or robots. LIDA is a conceptual and partly computational model that serves to implement and flesh out a number of psychological theories,[4] in particular the Global Workspace Theory of Baars.[5] Hence any LIDA agent, that is, any agent whose control structure is based on the LIDA Model, is at least functionally conscious.[6] Research on LIDA has entered its second decade.[7] This note is intended to summarize some of the newer developments of the LIDA Model.

## THE LIDA TUTORIAL

The LIDA Model is quite complex consisting of numerous independently and asynchronously operating modules (see Figure 1). It has been described in more than fifty published papers, presenting a considerable challenge to any would-be student of the model. Thus the recent appearance of a LIDA tutorial paper summarizing the contents of these earlier papers, as well as new material, is a significant new LIDA development.[8] The tutorial reduces the fifty some-odd papers into only fifty some-odd pages of text and figures.

## AI: ITS NATURE AND FUTURE

In 2016, Oxford University Press published philosopher/cognitive scientist Margaret Boden's *AI: Its Nature and Future*, which pays considerable attention to our LIDA Model.

Pointing out that LIDA "arises from a unified, systems-level theory of cognition," Boden goes on to speak of LIDA as being "deeply informed by cognitive psychology, having been developed for scientific, not technological, purposes," and "designed to take into account a wide variety of well-known psychological phenomena, and a wide range of experimental evidence." She says that "integrating highly diverse experimental evidence," LIDA is used "to explore theories in cognitive psychology and neuroscience." She also says that "the *philosophical* significance of LIDA, for instance, is that it specifies an organized set of virtual machines that shows how the diverse aspects of (functional) consciousness are possible." And Boden points out that the LIDA Model speaks to the "binding" problem, to the frame problem, and avoids any central executive.[9]

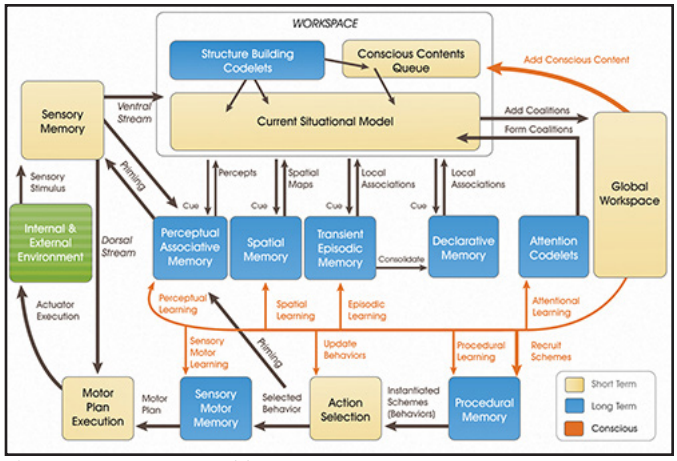


**Figure 1.** The LIDA Cognitive Cycle.

## ACTION EXECUTION

The LIDA Model attempts to model minds generally, providing an architecture for the control structure of any number of different LIDA-based agents. Thus, the LIDA Model in its general form must remain uncommitted to particular mechanisms or specifications for senses, actions, and environments. Each of its many independent and asynchronous modules, mentioned above, must allow for implementation so as to serve various agents with a variety of senses, actions, and environments.

Two of LIDA's most recently developed modules are devoted to action execution, which is concerned with creating a motor plan for a selected goal-directed behavior, and executing it. A motor plan template transforms a selected behavior into a sequence of executable actions. The Sensory Motor Memory (see Figure 1 above) learns and remembers motor plan templates.[10] Based on the subsumption architecture,[11] our LIDA agent testing this module adds analogs of the visual system's dorsal and ventral streams to the model. Given an appropriate motor plan for the selected behavior, the Motor Plan Execution module instantiates a suitable motor plan and executes it.[12] Together the two modules allow a LIDA-based agent to execute a selected action, quite important for any autonomous agent.

We have also introduced a new type of sensorimotor learning to the LIDA Model.[13] Using reinforcement learning, it stores and updates the rewards of pairs of data, motor commands, and their contexts, allowing the agent to output effective commands based on its reward history. As is all learning in LIDA, this sensorimotor learning is cued by the agent's conscious content. A dynamic learning

rate controls the effect of the newly arriving reward. The mechanism controlling the learning rate is inspired by the memory of errors hypothesis from neuroscience.[14] Our computer simulations indicate that using such a dynamic learning rate improves movement performance.

## SPATIAL MEMORY

In any cognitive system, memory is most generally defined as the encoding, storing, and recovery of information of some sort. The storage can be over various time scales. Cognitive modelers, and cognitive scientists in general, tend to divide the memory pie in many different ways. The LIDA Model has separate, asynchronous modules for memory systems of diverse informational types. (In Figure 1, the modules for various long-term memory systems are dark colored.) Much earlier research was devoted to Perceptual Associative Memory, Transient Episodic Memory, Declarative Memory, and Procedural Memory. (In all these cases, there is much left to be done.) Recent work on Sensory Motor Memory was discussed in the preceding section.

Over the past couple of years we have begun to think seriously about how best to represent data in Spatial Memory, representations of spatial information concerning objects in the agent's environment, and its location within it. We picture long-term Spatial Memory as consisting of hierarchies of cognitive maps, each representing the size, shape, and location of objects, and the directions and distances between them. In addition to long-term spatial memory, LIDA's working memory may contain one or a few cognitive map segments and facilitate planning and updating. Inspired by place and grid cells involved in spatial representations in mammalian brains, cognitive map representations in LIDA also consist of hierarchical grids of place nodes, which can be associated with percepts and events. We have implemented prototype mechanisms for probabilistic cue integration and error correction to mitigate the problems associated with accumulating errors from noisy sensors (see the section on uncertainty below). So far we have only experimented with how human agents mentally represent such cognitive maps of neighborhoods.[15]

## MOTIVATION

Every autonomous agent, be it human, animal, or artificial, must act in pursuit of its own agenda.[16] To produce that agenda requires motivation. Actions in the LIDA Model are motivated by feelings, including emotions—that is, feelings with cognitive content.[17] An early paper lays this out and relates feelings in this context to both values and utility.[18] More recent work fleshes out just how feelings play a major role in motivating the choice of actions.[19] Feelings arise in Sensory Memory (see Figure 1), are recognized in Perceptual Associative Memory, and become part of the percept that updates the Current Situational Model. There they arouse structure building codelets to produce various options advocating possible responses to the feeling, in accordance with appraisal theories of emotion.[20] The most salient of these wins the competition for consciousness in the Global Workspace and is broadcast, in particular, to Procedural Memory. There schemes proposing specific actions to implement the broadcast option are instantiated and forwarded to Action Selection, where a single action is selected as a response to the original feeling. Thus, feelings act as motivators.

## SELF

Any systems-level cognitive model such as our LIDA Model that aspires to model consciousness must attempt to account for the notion of self with its multiple aspects. We have made one attempt at describing how a number of different "selves" could be constructed within the LIDA Model.[21] These include the minimal (or core) self with its three sub-selves, self as subject, self as experiencer, and self as agent. The sub-selves of the extended self are comprised of the autobiographical self, the self-concept, the volitional (or executive) self, and the narrative self.

More recently we have begun to augment this account by combining these constructs with key elements of Shaun Gallagher's pattern theory of self, namely, his meta-theoretical list of aspects.[22] These include minimal embodied aspects, minimal experiential aspects, affective aspects, intersubjective aspects, psychological/cognitive aspects, narrative aspects, extended aspects, and situated aspects. We explore the use of the various aspects of this pattern theory of self in producing each of the various selves within the LIDA Model. The three types of minimal self are all implemented using only minimal embodied aspects and minimal experiential aspects. All of these can be created within the current LIDA Model. The four types of extended self will require all eight aspects in the list. Some of these will require additional processes to be added to the LIDA Model.

This use of pattern theory is helping us to clarify various theoretical issues with including various "selves" in the LIDA Model, as well as open questions such as the relationships between different sub-selves. Using pattern theory also can enable us to set benchmarks for testing for the presence of various types of self in different LIDA-based agents.

## CYCLIC TO MULTICYCLIC PROCESSES

The LIDA Model begins its fleshing out of Global Workspace Theory by postulating a cognitive cycle (see Figure 1 for a detailed diagram), which we view as a cognitive atom from which more complex cognitive processes are constructed. A LIDA agent spends its "life" in a continual, cascading (overlapping) sequence of such cognitive cycles, each sensing and understanding the agent's current situation, and choosing and executing an appropriate response. Such cycles occur five to ten times a second in humans.[23] The first decade or more of our research was devoted to trying to understand what happens during a single cognitive cycle, taking in humans 200 to 500 ms. Now, having at least a partial overall discernment of the activity of a single cycle, we feel emboldened to turn some of our attention to more complex multi-cyclic processes such as planning, reasoning, and deliberation.

## LANGUAGE

LIDA has been criticized for focusing on low intelligence tasks, and lacking high cognitive functions such as language understanding.[24] To overcome this gap, and initiate language processing in the LIDA architecture,

learning the meaning of the vervet monkey alarm calls was simulated. Field studies revealed the existence of three distinct alarm calls.[25] Each call is emitted to warn the rest of the group of the danger from a predator in the vicinity. Upon hearing a particular alarm call, vervet monkeys typically escape into safe locations in a manner appropriate to the predator type signaled by that alarm. A LIDA-based agent that learns the meaning of these alarm calls has been developed.[26] LIDA's perceptual learning mechanism was implemented to associate each alarm call with three distinct meanings: an action-based meaning, a feeling-based meaning, and a referential-based meaning. This multiple-meaning-assessment approach aligns with our ultimate goal of modeling human words that must convey multiple meanings. A manuscript describing this research has been submitted, reviewed, revised, and resubmitted.[27]

## LIDA'S HYPOTHESIS REGARDING BRAIN RHYTHMS

Marr proposed three levels of analysis for cognitive modeling—the computational, the representational/algorithmic, and the implementational.[28] As a general model of minds, LIDA's core concepts possess an applicability that spans many possible domains and implementations. Accordingly, LIDA's primary area of interest lies within Marr's computational and algorithmic levels. However, many classes of biological mind fall within LIDA's purview, and modeling biological minds from the perspective of the LIDA Model requires careful attention to the available evidence and the competing theories regarding the way in which brains affect control structures for behavior in humans and certain non-human animals.

A helpful metaphor may be found in the example problem of reverse engineering a software program. The primary goal is to uncover the algorithms that carry out the software's computations, but this might require, or at least be facilitated by, investigation of the operations carried out in the hardware during the program's execution. We frequently assert that LIDA is a model of minds rather than brains. Having said that, we find that understanding those biological minds of interest to LIDA requires close and frequent reference to the way brains carry out computations. In practice, this has meant examination of biological minds at the implementation level as well as the algorithmic and computational levels.

While neuroscience manifests a solid theoretical consensus regarding the basic tenets of neuroanatomy and neuronal physiology, considerable controversy continues to pervade investigations into the cognitive aspects of neural function. The vast proliferation of evidence resulting from recent decades' technological advances have thus far failed to converge on a consensual framework for understanding the neural basis of cognition. Nonetheless, LIDA's perspective on biological minds currently commits to a particular collection of theoretical proposals situated squarely within the broader controversy. While a detailed treatment of these proposals lies outside the scope of the present discussion, we give a brief overview as follows.

The Cognitive Cycle Hypothesis and the Global Workspace Theory (GWT) of Consciousness form the backbone of the LIDA Model. GWT, originally a psychological theory,[29] was recently updated into a neuropsychological theory known as Dynamic Global Workspace Theory (dGWT).[30] Per dGWT, a global workspace is "a dynamic capacity for binding and propagation of neural signals over multiple task-related networks, a kind of neuronal cloud computing."[31] Per LIDA's Cognitive Cycle Hypothesis, the global workspace produces a quasiperiodic broadcast of unitary and internally consistent cognitive content that mediates an autonomous agent's action selection and learning, and, over time, comprises the agent's stream of consciousness.

The theoretical proposals of Freeman's Neurodynamics provide the framework most harmonious with LIDA's central hypotheses.[32] Within this framework, a cognitive cycle comprises the emergence of a self-organized pattern of neurodynamic activity. LIDA's Rhythms Hypothesis postulates that the content of a cycle's broadcast from the global workspace manifests in experimentally observable brain rhythms as gamma (30-80 Hz) frequency activity scaffolded within a slow-wave structure of approximately theta (4-6 Hz) frequency that tracks the rhythm of successive broadcasts. Elaboration of this hypothesis within the framework of Freeman's neurodynamical theory is quite complex and is the subject of a publication currently under preparation.

## MENTAL IMAGERY, PRECONSCIOUS SIMULATION, AND GROUNDED COGNITION

Most humans report the ability to have sensory-like experiences in the absence of external stimuli. They describe experiences such as "having a song stuck in our heads" or "listening to our inner voices" or "seeing with our mind's eye." In the literature cited below, these phenomena are referred to as "mental imagery." Many experiments have been performed that suggest mental imagery facilitates, and may be critical for, a broad range of mental activities including prediction,[33] problem solving,[34] mental rehearsal,[35] and language comprehension.[36] Cognitive models are needed to help explain the processes that underlie mental imagery. We have begun to leverage the LIDA model to gain insight into how the fundamental capabilities needed for mental imagery could be realized in artificial minds and to apply these insights toward the construction of software agents that utilize mental imagery to their advantage.

Mental imagery is by definition a conscious process; however, there is an intriguing possibility that the same mechanisms underlying mental imagery also support preconscious cognitive processes and enable grounded (embodied) cognition. The psychologist and cognitive scientist Lawrence Barsalou defines "simulation" as the "re-enactment of perceptual, motor, and introspective states acquired during experience with the world, body, and mind," and hypothesizes that

> [simulation] is not necessarily conscious but may also be unconscious, probably being unconscious even more often than conscious.

> Unconscious [simulations] may occur frequently during perception, memory, conceptualization, comprehension and reasoning, along with conscious [simulations]. When [simulations] reach awareness, they can be viewed as constituting mental imagery.[37]

It is a goal of our research program to explore the possibility of a unified set of mechanisms supporting mental imagery, preconscious simulation, and grounded cognition. The LIDA Model provides an ideal foundation for exploring these topics, as it is one of the few biologically inspired cognitive architectures that attempts to model functional consciousness, and is firmly committed to grounded cognition.[38]

## REPRESENTING AND COMPUTING WITH UNCERTAINTY IN LIDA

Cognition must deal with large amounts of uncertainty due to a partially observable environment, erroneous sensors, noisy neural computation, and limited cognitive resources. There is increasing evidence for probabilistic mechanisms in brains.[39] We have recently started exploring probabilistic computation for LIDA, as of now for the specific purpose of dealing with spatial uncertainty and complexity in navigation.[40] Work is underway to augment LIDA's representations (inspired by Barsalou's perceptual symbols and simulators[41]) with a representation and computation mechanism accounting both for the uncertainty in various domains as well as approximately optimal inference given cognitive, time, and memory limitations.

## LIDA FRAMEWORK IN PYTHON

In 2011, Snaider et al. presented the "LIDA Framework," a software framework written in the Java programming language that aims to simplify the process of developing LIDA agents.[42] The LIDA Framework implements much of the low-level functionality that is needed to create a LIDA software agent and provides default implementations for many of the LIDA modules. As a result, simple agents can often be created with a modest level of effort by leveraging "out of the box" functionality.

Inspired by the success of the LIDA Framework, a sister project is underway to implement a software framework in the Python programming language, which we refer to as lidapy. One of lidapy's primary goals has been to facilitate the creation of LIDA agents that are situated in complex and "real world" environments, with the eventual goal of supporting LIDA agents in a robotics context. Toward this end, lidapy has been designed from the ground up to integrate with the Robot Operating System, a framework developed by the Open Source Robotics Foundation (OSRF) that was specifically designed to support large-scale software development in the robotics domain.[43]

### NOTES

1. S. Franklin, *Artificial Minds* (Cambridge, MA: MIT Press, 1995), 412.
2. S. Franklin and A. C. Graesser, "Is It an Agent, or Just a Program?: A Taxonomy for Autonomous Agents," *Intelligent Agents III* (Berlin: Springer Verlag, 1997), 21–35.
3. For historical reasons, this word was previously "distribution." It has been recently changed to better capture important aspects of the model in its name.
4. A. D. Baddeley, "Working Memory and Conscious Awareness," in *Theories of Memory*, ed. A. Collins, S. Gathercole, Martin A. Conway, and P. Morris, 11–28 (Howe: Erlbaum, 1993); L. W. Barsalou, "Perceptual Symbol Systems," *Behavioral and Brain Sciences* 22 (1999): 577–609; Martin A. Conway, "Sensory–Perceptual Episodic Memory and Its Context: Autobiographical Memory," *Philos. Trans. R. Soc. Lond B.* 356 (2001): 1375–84; K. A. Ericsson and W. Kintsch, "Long-Term Working Memory," *Psychological Review* 102 (1995): 211–45; A. M. Glenberg, "What Memory Is For," *Behavioral and Brain Sciences*, 20 (1997): 1–19; M. Minsky, *The Society of Mind* (New York: Simon and Schuster, 1985); A. Sloman, "What Sort of Architecture Is Required for a Human-Like Agent?" in *Foundations of Rational Agency*, ed. M. Wooldridge and A. S. Rao, 35–52 (Dordrecht, Netherlands: Kluwer Academic Publishers, 1999).
5. Bernard J. Baars, *A Cognitive Theory of Consciousness* (Cambridge: Cambridge University Press, 1988).
6. S. Franklin, "IDA: A Conscious Artifact?" *Journal of Consciousness Studies* 10 (2003): 47–66.
7. S. Franklin and F. G. J. Patterson, "The LIDA Architecture: Adding New Modes of Learning to an Intelligent, Autonomous, Software Agent," *IDPT-2006 Proceedings (Integrated Design and Process Technology)*: Society for Design and Process Science, 2006.
8. S. Franklin, T. Madl, S. Strain, U. Faghihi, D. Dong, et al., "A LIDA Cognitive Model Tutorial," *Biologically Inspired Cognitive Architectures* (2016): 105–30. doi: 10.1016/j.bica.2016.04.003.
9. M. A. Boden, *AI: Its Nature and Future* (Oxford, UK: Oxford University Press, 2016), 98–128.
10. D. Dong and S. Franklin, "Sensory Motor System: Modeling the Process of Action Execution," paper presented at the Proceedings of the 36th Annual Conference of the Cognitive Science Society, 2014.
11. R. Brooks, "A Robust Layered Control System for a Mobile Robot," *IEEE Journal of Robotics and Automation* 2, no. 1 (1986): 14–23.
12. D. Dong and S. Franklin, "A New Action Execution Module for the Learning Intelligent Distribution Agent (LIDA): The Sensory Motor System," *Cognitive Computation* (2015). doi: 10.1007/s12559-015-9322-3.
13. D. Dong and S. Franklin, "Modeling Sensorimotor Learning in LIDA Using a Dynamic Learning Rate," *Biologically Inspired Cognitive Architectures* 14 (2015): 1–9.
14. D. J. Herzfeld, P. A. Vaswani, M. K. Marko, and R. Shadmehr, "A Memory of Errors in Sensorimotor Learning," *Science* 345, no. 6202 (2014): 1349–53.
15. Tamas Madl, Stan Franklin, Ke Chen, Daniela Montaldi, and Robert Trappl, "Towards Real-World Capable Spatial Memory in the LIDA Cognitive Literature," *Biologically Inspired Cognitive Architectures* 16 (2016): 87–104; Tamas Madl, Stan Franklin, Ke Chen, Robert Trappl, and Daniela Montaldi, "Exploring the Structure of Spatial Representations," *PLoS ONE* 11, no. 6 (2016): e0157343.
16. Franklin and Graesser, "Is It an Agent, or Just a Program?"
17. Victor S. Johnston, *Why We Feel:The Science of Human Emotions* (Reading MA: Perseus Books, 1999).
18. S. Franklin and U. Ramamurthy, "Motivations, Values, and Emotions: Three Sides of the Same Coin," *Proceedings of the Sixth International Workshop on Epigenetic Robotics*, Vol. 128 (Paris, France: Lund University Cognitive Studies, 2006), 41–48.
19. R. McCall, *Fundamental Motivation and Perception for a Systems-Level Cognitive Architecture*, PhD Thesis, University of Memphis, Memphis, TN, USA, 2014; R. J. McCall, S. Franklin, U. Faghihi, and J. Snaider, "Artificial Motivation for Cognitive Software Agents," submitted.
20. Franklin et al., "A LIDA Cognitive Model Tutorial."
21. U. Ramamurthy and S. Franklin, "Self System in a Model of Cognition," paper presented at the Machine Consciousness Symposium at the Artificial Intelligence and Simulation of Behavior Convention (AISB'11), University of York, UK, 2011.

22. S. Gallagher, "A Pattern Theory of Self," *Frontiers in Human Neuroscience* 7, no. 443 (2013): 1–7.

23. T. Madl, B. J. Baars, and S. Franklin, "The Timing of the Cognitive Cycle," *PLoS ONE* 6, no. 4 (2011): e14803. doi: 10.1371/journal.pone.0014803.

24. W. Duch, R. Oentaryo, and M. Pasquier, "Cognitive Architectures: Where Do We Go From Here?" in *Artificial General Intelligence, 2008: Proceedings of the First AGI Conference*, ed. P. Wang, B. Goertzel, and S. Franklin, 122–37 (2008).

25. R. Seyfarth, D. Cheney, and P. Marler, "Monkey Responses to Three Different Alarm Calls: Evidence of Predator Classification and Semantic Communication," *Science* 210, no. 4471 (1980): 801–03.

26. N. A. Khayi-Enyinda, "Learning the Meaning of the Vervet Alarm Calls Using a Cognitive and Computational Model," Master of Science, University of Memphis, 2013.

27. N. Ait Khayi and S. Franklin, "Initiating Language in LIDA: Learning the Meaning of Vervet Alarm Calls," *Biologically Inspired Cognitive Architectures* 23 (2018): 7–18. doi: 10.1016/j.bica.2018.01.003.

28. D. C. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (New York: Freeman, 1982).

29. Baars,*A Cognitive Theory of Consciousness*.

30. B. Baars, S. Franklin, and T. Ramsøy, "Global Workspace Dynamics: Cortical 'Binding and Propagation' Enables Conscious Contents," *Frontiers in Consciousness Research* 4, no. 200 (2013). doi: 10.3389/fpsyg.2013.00200.

31. Baars et al., "Global Workspace Dynamics," 1.

32. W. Freeman, *Neurodynamics: An Exploration in Mesoscopic Brain Dynamics* (Springer Science & Business Media, 2012); W. J. Freeman and R. Kozma, "Freeman's Mass Action," *Scholarpedia* 5, no. 1 (2010): 8040.

33. S. T. Moulton and S. M. Kosslyn, "Imagining Predictions: Mental Imagery as Mental Emulation," *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 364, no. 1521 (2009): 1273–80.

34. Y. Qin and H. A. Simon, "Imagery and Mental Models in Problem Solving," paper presented at the Proc. AAAI Symposium on Reasoning with Diagrammatic Representations, Stanford, CA, 1992; P. Shaver, L. Pierson, and S. Lang, "Converging Evidence for the Functional Significance of Imagery in Problem Solving," *Cognition* 3, no. 4 (1975): 359–75.

35. J. E. Driskell, C. Copper, and A. Moran, "Does Mental Practice Enhance Performance?" American Psychological Association, 1994; P. E. Keller, "Mental Imagery in Music Performance: Underlying Mechanisms and Potential Benefits," *Annals of the New York Academy of Sciences* 1252, no. 1 (2012): 206–13.

36. B. K. Bergen, S. Lindsay, T. Matlock, and S. Narayanan, "Spatial and Linguistic Aspects of Visual Imagery in Sentence Comprehension," *Cognitive Science* 31, no. 5 (2007): 733–64; R. A. Zwaan, R. A. Stanfield, and R. H. Yaxley, "Language Comprehenders Mentally Represent the Shapes of Objects," *Psychological Science* 13, no. 2 (2002): 168–71.

37. L. W. Barsalou, "Simulation, Situated Conceptualization, and Prediction," *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 364, no. 1521 (2009): 1281–89.

38. S. Franklin, S. Strain, R. McCall, and B. Baars, "Conceptual Commitments of the LIDA Model of Cognition," *Journal of Artificial General Intelligence* 4, n. 2 (2013): 1–22. doi: 10.2478/jagi-2013-0002.

39. N. Chater, J. B. Tenenbaum, and A. Yuille, "Probabilistic Models of Cognition: Conceptual Foundations," *Trends in Cognitive Sciences* 10, no. 7 (2006): 287–91; A. Clark, "Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science," *Behavioral and Brain Sciences* 36, no. 03 (2013): 181–204; D. C. Knill and A. Pouget, "The Bayesian Brain: The Role of Uncertainty in Neural Coding and Computation," *TRENDS in Neurosciences* 27, no. 12 (2004): 712–19.

40. T. Madl, S. Franklin, K. Chen, R. Trappl, and D. Montaldi, "Exploring the Structure of Spatial Representations," *PLoS ONE* (2016); T. Madl, S. Franklin, K. Chen, D. Montaldi, and R. Trappl, "Towards Real-World Capable Spatial Memory in the LIDA Cognitive Architecture," *Biologically Inspired Cognitive Architectures* 16 (2016): 87–104, doi: 10.1016/j.bica.2016.02.001.

41. Barsalou, "Perceptual Symbol Systems."

42. J. Snaider, R. McCall, and S. Franklin, "The LIDA Framework as a General Tool for AGI," paper presented at the Artificial General Intelligence (AGI-11), Mountain View, CA, 2011.

43. M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, et al., "ROS: An Open-Source Robot Operating System," paper presented at the ICRA workshop on open source software, 2009.

## *Distraction and Prioritization: Combining Models to Create Reactive Robots*

Jonathan R. Milton
**UNIVERSITY OF ILLINOIS SPRINGFIELD**

In this paper, I intend to present a theoretical framework for combining existing cognitive architectures, in order to fully and specifically address the areas of distraction and prioritization in autonomous systems. The topic of this paper directly addresses an issue which was raised by Troy Kelley and Vladislav Veksler in their paper "Sleep, Boredom, and Distraction: What Are the Computational Benefits for Cognition?"[1] Specifically, I intend to focus mainly on the theme of "distraction" with regard to their paper, as that is the area Kelley and Veksler seemed to have the most difficulties with, regarding the compatibility of various design options.

As researchers at the US Army Research Laboratory, Kelly and Veksler are trying to create a robot that has the ability to prioritize goals in consistently unpredictable environments. In their paper, Kelley and Veksler show how the ability to become distracted turns out to be a critical component of how humans prioritize their goals. Kelley and Veksler would like their robot to be able to be appropriately distracted from any initial prime mission focus whenever urgent and unexpected changes occur within the robot's operational environment. Their argument on behalf of distraction, along with their stated goals, has led me to explore possible cognitive structures that could allow for task-specific concentrations to be combined with outside world information processing, in order to allow for effective goal prioritization. I intend to show that task-specific concentrations can be instilled through procedural learning and habituation, while simultaneous outside world information processing can occur with the added help of specially installed processors. The intent is that these special processors will operate in a manner that appears to mimic the seemingly innate abilities in humans, which often assist us with intuitively predicting physical reactions, as well as with identifying potentially dangerous situations.

As with other cognitive-science-related fields, the study of artificial intelligence regularly involves an interdisciplinary approach in conjunction with philosophy. The main topics discussed in this paper, as they relate to philosophy, are the areas of artificial emotions and innate knowledge. This paper undoubtedly takes a cognitive appraisal view

of emotions in that emotional experiences in machines are probably best described as being determined by the evaluation of a certain stimulus.[2] Beliefs, desires, and judgments are generally not involved in the descriptions of emotional states involving machines. The emphasis regarding emotional content in machines is usually focused on processes and perceptions, as opposed to the subjective experience of a biologically produced emotional state. The cognitive appraisal view of emotions is widely accepted in both the fields of psychology and philosophy, and while debate certainly still exists on the matter (mainly involving propositional attitudes), I do not anticipate too many objections to the strict adherence to the cognitive appraisal view in this instance. Furthermore, this paper undoubtedly assumes that innate knowledge is an indispensable feature for developing the superior cognitive abilities found in humans. While reliable research exists to add weight to the claim of humans having at least *some* form of innate knowledge, I do not intend to present an argument for that particular position. Rather, the focus on innate knowledge in this paper is to show how it could be used as an invaluable shortcut for giving autonomous machines certain abilities, based on the needs of their particular function.

The goal of this paper is to show that existing models could hypothetically be combined into one autonomous machine which would allow for distractibility and adaptive prioritization. For the sake of providing some direction to this design project, let us say that our hypothetical robot (who we'll call PARS, Priority-based Adaptive Reaction System) is to be a combat robot designed for protecting buildings and rooms, as in the example provided by Kelley and Veksler.

To accomplish the goals outlined above, I intend to draw attention to models such as LIDA,[3] Argus Prime,[4] and IPE,[5] in order to show how elements of these three systems can be combined to produce a model that more specifically suits the hypothetical robot design for the purposes outlined below. My focus as far as inspiration from the field of neuroscience will, like the LIDA model, rely heavily on Bernard Baars's *global workspace theory* (GWT).

## WHY IS DISTRACTION IMPORTANT?

People may not realize that distraction actually plays a vitally important role in how priorities and goal selections are created. Humans get mentally distracted sometimes without consciously realizing it, and as Kelley and Veksler point out in their paper, goal forgetting actually occurs when an agent's focus of attention shifts, due to either external cues or tangential lines of thought. Without distraction, humans could potentially begin a task—for whatever reason—and that task would become their all-consuming priority regardless of its importance. Furthermore, the task in question would remain a person's sole focus until it was completely finished. If a person's goal was to clean up their bedroom, then they would clean their bedroom until their task was complete, ostensibly even if their house was engulfed in flames around them.

As Kelley and Veksler also address in their paper, "novelty" is a highly important feature for redirecting attention, when needed, and consistently serves to prevent boredom. Furthermore, stressful situations can create a sense of urgency and lessen the chances of one being distracted through a phenomenon known as "cognitive tunneling." As will be discussed later in this paper, less stressful situations can create a more comfortable and largely predictable environment, which would allow for the natural emphasizing of contrasts.

At first glance, distractedness seems to be a suboptimal and inefficient aspect of human cognition; however, as Kelley and Veksler have correctly pointed out, being able to be distracted and thus adjust one's priorities turns out to be a critically important feature of human consciousness.

## TRANSFERENCE TO ROBOTS

Since emphasis has now been placed on the importance of distraction for human operations and activities, we should naturally be able to see how that same feature can be beneficial for any machines that humans may attempt to design and ultimately entrust with extremely important responsibilities. There seems to be some difficulty, however, when it comes to actually giving machines this crucial ability. The difficulty appears to lie in assigning specific tasks to robots, yet also giving these robots the ability to adjust their priorities whenever necessary. In other words, how do we tell a machine to do one task, yet allow that machine to become distracted and select a different, yet appropriate, task/goal without specifically commanding the robot to do so? As stated above, the goal of this paper is to try and design a robot model that could allow for necessary distractedness, and then ultimately achieve effective goal prioritization.

## INNATE ABILITIES

I would like to begin the design process by focusing on the topic of innate abilities. The topic of innate abilities in humans has been studied and debated for centuries, and rather than revisit those debates here, my aim is to draw particular attention to the seemingly innate knowledge of *physical reasoning* and *physical scene understanding* in humans. Believe it or not, infants as young as two months old display a basic understanding that physical laws exist, as well as an expectation that those laws will always be obeyed. Research being conducted by top contemporary psychologists show that physical scene understandings appear in humans at such an early age that it gives the appearance of humans possessing innate concepts and specialized learning mechanisms.[6] It would seem almost like a natural conclusion that the most effective way to create a machine that is capable of mimicking the human cognitive abilities of being distracted, assessing situations, prioritizing goals, etc., would be to try and recreate the functional processes by which humans acquire those abilities in the first place. If innate abilities appear to be a fundamental aspect of human cognition, then why should we not try and come up with a design that could seemingly imitate that process in intelligent machines?

## SPATIOTEMPORAL EMPHASIS

An additional important topic worth discussing is placing an emphasis on spatiotemporal processing as being a critical aspect of early developmental learning in machines.

Most machine-learning literature I have researched tended to focus mainly on feature detection for object recognition, while spatiotemporal awareness appears to be viewed as an assumed consequence of robots interacting with their environments. While there is a great deal of focus and research dedicated to spatial-temporal processing in machine vision, there seems to be a persistence of emphasizing—or natural relying upon—feature detection as being the most vital component of identifying objects.

In "Objects and Attention: The State of the Art," Brian Scholl writes how spatiotemporal features could be more "tightly coupled" with object representations than surface-based features such as "color and shape." In fact, when it comes to human development, Scholl highlights studies, that show how ten-month-old infants will use spatiotemporal information, but not featural information, in order to assess an object's unity.[7] Scholl further explains that typically, once an infant reaches twelve months, studies then show that the infant will begin to use both spatiotemporal and featural information processing for object recognition, which then becomes the persistent interactive object recognition process that carries into adulthood.

All of that said, it seems that a more natural development of machine vision/intelligence systems should approach training robots by first focusing on spatiotemporal information processing, and then moving on to using an interaction-type process of both spatiotemporal and feature-detection processing for object recognition. In my opinion, this ideal achievement would be critical for the successful operation of PARS in the developmental stage, especially when the goal is to then install existing models to be used to mimic the "special innate processes" that are so vital to the way humans analyze the world around them.

## BACKGROUND ON MODEL EXAMPLES USED

Turning attention back to our hypothetical robot design, after a basic developmental stage (focusing first on spatiotemporal processing, as outlined above), I would like to address the specific models that could be used to give PARS the seemingly innate abilities of humans, which can then be used to assist with accomplishing specific tasks, while also allowing for distraction. I will briefly state—and then outline below—that I believe a pre-programmed *intuitive physics engine* (or IPE) and an object motion classification processor such as the Argus Prime could potentially help PARS to perform procedural tasks faster by identifying items more quickly, and ultimately select goals more efficiently after a distracted period. Furthermore, the most important operational model is the LIDA, as it would serve as the foundational model that the other two aforementioned models would be used in conjunction with.

### 1) LIDA

The LIDA model was designed at the University of Memphis under the direction of Stan Franklin. The LIDA team draws inspiration from Bernard Baars's *global workspace theory* by creating a coalition of small pieces of independent codes called *codelets* (or sometimes referred to as "processors"). These codelets search out items that interest them—such as novel or problematic situations—which can then be broadcast as vital messages to the entire network of processors in order to recruit enough internal resources to handle a particular situation.[8] The LIDA seems like an ideal scheme for my intentions, and I will draw on this model quite heavily. I intend to rely on specific areas of the LIDA such as its ability to do the following:

a) Use episodic memory for long-term storage of autobiographical and semantic information

b) Use its serial yet overlapping cognitive cycles to facilitate perception, local associations (based off of memories and emotional content), codelet competition (used for locating novel or urgent events), conscious broadcasting (the network recruitment of processors to handle novel/ urgent events), setting goal context hierarchy, and, finally, selecting and taking appropriate action.

### 2) Argus Prime

The Argus Prime model was designed at George Mason University by Michael Schoelles and Wayne Gray for the purpose of operating in a complex simulated task environment. Argus Prime is tasked with performing functions similar to a human radar operator. Argus Prime must complete subtasks such as identifying, classifying, and reacting to targets/threats. Argus Prime is based off of the ACT-R/PM process of parallel elements of cognition, perception, and motor movement.

### 3) Intuitive Physics Engine (IPE)

This model was outlined by research scientists at the Brain and Cognitive Sciences Department at Massachusetts Institute of Technology and should probably, and more accurately, be called the Open Dynamics Engine used in conjunction with a Bayesian Monte Carlo simulation approach. The intent of this model is actually to *mimic* the human IPE that most accurately describes how we use our understanding of "geometries, arrangements, masses, elasticities, rigidities, surface characteristics and velocities" to predict probable outcomes in complex natural scenes.[9]

## LIDA AND THE COGNITIVE CYCLE

Before describing how these models could be combined to suit PARS's operational needs, I would like to first outline exactly how these models could theoretically fit together in the design stage.

The LIDA model is highly complex, and it should be stated upfront that in order to fully understand how this model functions, one really should take the time to read Stan Franklin and Co.'s description of it (see references). For my purposes, I will present only an abbreviated description of LIDA's cognitive cycle, in addition to the basic operational features outlined above. The serial process of LIDA's cognition cycle begins with an external stimulus which travels through specific modules for certain purposes, such as the *perceptual associative memory* module for category representation; the *workspace* module for creating the temporary structures which are used to potentially distribute information to the requisite processors; the

*episodic, declarative, and procedural memories* modules for different storage and use purposes; and, lastly, an *action selection* module. Reasoning and problem-solving occur over multiple cognitive cycles in the LIDA model, and included in those multicyclic processes are the features of *deliberation*, *voluntary action*, *non-routine problem-solving*, and *automatization*.[10]

Given that LIDA relies on a coalition of special processors to work together for a specific task, then it seems quite feasible that additional space could be made for the insertion of processors containing specifically constructed subsets of data, in order to create the predisposition in PARS towards a particular approach when conducting outside world information processing. This ingrained approach would be the quality that gives PARS the appearance of having innate attributes, as the tendency towards that particular approach would not be the result of a "learned process."

Since we can now feasibly include additional processors into the pre-existing LIDA design, then why not seek out existing models to serve as the specially added processors which can address the areas needed for PARS's specific purpose of function? Enter the IPE and AP models for physical scene understanding and threat classification, respectively. Threat classification and physical scene understanding should naturally stand out as two critical and necessary abilities required for any agent tasked with providing physical security. This is because visually acquiring and identifying potential threats is probably the most important task required of a security agent. Furthermore, any potential action/physical response by a security agent that has identified a threat would need to undergo an analysis of what can and cannot be physically done in that particular operational environment (more on this later).

Given that the two features outlined above are so critical to the specific operations of PARS, it seems quite reasonable that the IPE and AP models would be better emphasized as their own *modules* or *sub-modules* within the actual LIDA cognitive cycle. This would allow these vital modules to work directly with the *workspace* module on a constant basis. For example, the IPE and AP classifier could be placed alongside the *transient episodic memory* module and the *declarative memory* module in the existing LIDA model diagram (see Figure 1); or they could potentially fit as automatically involved sub-modules alongside the *structure building* and *attention codelet* modules. Either way, the intent would be for both of those critical areas to be visited mandatorily once every cognitive cycle, which already happens at around once every 380ms.[11]

At this point, it seems necessary to draw attention to the actual data content that will be present in the AP and IPE models/modules that will be used in PARS. The IPE model seems perfectly suited as it is, for our purposes, and a special processor with just the data required for a functioning IPE can be installed as is, on top of the current LIDA model, with communication pathways linked between the IPE module and the LIDA workspace module (see lower left portion of Figure 1).

The AP-styled model/module would operate similar to the IPE, and contain pre-programmed data which could be installed onto the LIDA model. However, the data in the AP "like" model for our purposes would be somewhat different from the Argus Prime in that the threat element data in PARS would need to consist of a catalog of weapons and other potential threat components, as well as how those weapons and threat components normally function. This differs to a significant degree from the original AP model, which simply tries to determine the position and velocity of potential threats. The newly updated weapons data catalog for PARS will be accumulated and stored in this specific AP-like processor from the very first moment PARS becomes operational. Furthermore, the ACT-R/PM-based design of the AP model would seem to be an easily compatible processor for use within the larger LIDA operational design, as both models are serial-based systems that still allow for parallel information processing.[12]

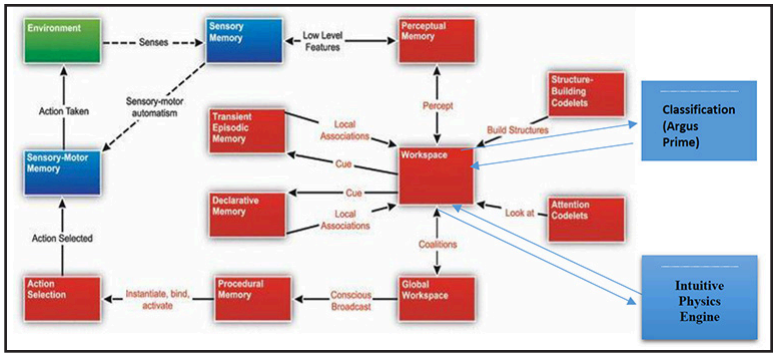


**Figure 1.** Current LIDA cognitive cycle diagram with added modules.

## DISTRACTION

Hopefully, at this point it is clear that

a) Distractibility is an important aspect of prioritization and goal selection.

b) Innate abilities appear necessary to mimic human cognitive abilities.

c) Feasible options exist to combine models in order to potentially achieve both *a* & *b* in autonomous machines.

Turning attention back to the issue of distractibility, I would like to present a detailed description of how the functional process of PARS would work to allow for distractedness and goal context hierarchy in a given operational environment. In order to better understand how PARS would become distracted, it might help to first analyze how it is that humans tend to become distracted.

Looking at the most common examples of what causes distraction in humans, I think most people would agree that unfamiliar objects and/or novel situations can create a sense of intrigue, which can lead to distracted mental states. This is especially true if those novel items/situations have the potential to become *emotional stressors*, by presenting a physical threat to an object or being that a person has conditioned a deep attachment toward. Humans always

seem to be on something like a subconscious standby mode, which is contingent on potential threats directed at things we value the most like our loved ones, personal safety, treasured belongings, etc. A threat toward any of those items (to name a few) would most likely trigger emotional stress and alter whatever priorities we may have held prior to noticing the potential threat. Therefore, emotional stress is an extremely effective way to create a distraction.

Another example of instances that create distractions in humans would be observing anything that offends our IPE (such as a floating table, or a person who walks through brick walls, etc.). Extraordinary physical anomalies will almost always turn our attention from one object/situation to another.

Lastly, humans tend to get comfortable with the familiar and the mundane. Whenever humans are repeatedly exposed to a particular stimulus, they will eventually start to have diminishing emotional reactions to that stimulus. In the field of psychology, this experience is referred to as *habituation*. If a person develops habituation within a certain environment, then encountering something new or unfamiliar within that environment will often grab a person's attention (to some degree), and normally distract said person away from any previously engaged activity.

The elements of habituation and facilitating emotional stress are where I think the GWT-structured LIDA system can be immensely beneficial for the function of PARS. Addressing the area of habituation first, the LIDA model's perceptual associated and episodic-oriented memory can be used to allow us to get PARS well accustomed to its operational environment via multiple walkthroughs. Furthermore, the LIDA model strives for automatization, which is ideal for the design of PARS in that procedural tasks (such as roaming/guarding a building perimeter) are learned to a point where they can be accomplished without constant conscious attention/focus. Operating successfully along those lines, any significant anomaly produced in PARS's operational environment would most likely be noticed, and therefore hopefully distract PARS's attention from its automatized task and initiate a potential threat-assessment sequence.

Whenever potentially distracting elements appear as noticeable irregularities within an operational environment, then those irregularities should serve as "cues" to initiate a process that puts elements of PARS's cognitive cycle on alert. This "alert" status of cognitive processing is where the LIDA design begins to recruit additional processors in order to determine how it will handle novel situations. The framework of commonly used cognitive processors is already functioning due to its conditioned use in the regular operational activities formed during the procedural learning process; however, additional processors can now be recruited in order to handle novel situations. Depending on the evaluation of any newly observed stimulus, these newly recruited processors may potentially produce an emotionally stressed state, allowing for intense focus via cognitive tunneling.

Similarly to what was outlined in the preceding paragraphs regarding habituation for perceptual familiarity, the LIDA model uses an "attachment period" to build emotional attachments. These attachments can also be used as primary motivators in the learning environment.[13] Emotional stressors could be things such as potential threats toward familiar building occupants that PARS is assigned to protect, as well as potential threats to sensitive objects and equipment that PARS has been conditioned to see as critically important. Any increased threats to those items would increase emotional stress in PARS, and potentially produce the cognitive tunneling that would block out any lesser important external information processing. It must be stated that the cognitive tunneling ability could have a potential downside to it and expose PARS to vulnerabilities when it comes to intentional deceptions. Admittedly, this is a challenge. Yet, it is no different than challenges that currently exist when humans become too narrowly focused on a given task/priority.

## PRIORITIZATION

Once PARS can notice environmental anomalies and emotional cues, then there is room to now advance on to the analysis phase and determine if any differences in the operational environment are worthy of PARS alternating its priorities from its primary task, which in this case would be to guard/patrol a specific route in an important building. It is worth explaining for the sake of clarification that a necessary feature of being "distracted" is *prioritization*, as one without the other would simply be a description of being aimless. An agent only becomes distracted when its attention has been drawn from one task or idea to another, and a distracted period only ends when an agent realizes the distraction and makes a goal selection in accordance with the agent's top priorities. Therefore, prioritization sequencing must be a necessity for anyone attempting to create effective distractibility in autonomous machines. The prioritization sequencing process used for PARS is approached by focusing on three specific goals:

1) Have PARS identify the most important danger (or potential catastrophe) in its environment by using a *classification* system that identifies threats and other dangerous situations.

2) Utilize a framework—much like a physics engine—that allows PARS to simultaneously observe and analyze large numbers of objects and events in order to *determine* the most likely *outcomes* of the observed situation.

3) Process all of the observations and analysis outlined in areas 1 and 2 by using the two additional models in conjunction with the LIDA cognitive cycle to facilitate deliberation in order to determine the following:

    a) Goal context hierarchy.

    b) Actions chosen/taken.

### *GOAL 1: THREAT CLASSIFICATION*

The Argus Prime (AP) model outlined above is able to recognize and analyze threats based on a variety of spatial and motion elements that must be taken into account, such as range, speed, course, and altitude. This is done in order to partly classify the threat level of the object that Argus Prime is observing/analyzing. For PARS's purposes, I would like to focus on specific threat classifications outlined and emphasized in advance through the "innate-like" inclusion of the AP-styled module/sub-module in the cognitive cycle portion.

Once PARS possesses a threat classification system for both motion (speed, range, vector, etc.) as well as for spatial residence (i.e., the exact spatial location the threating agent occupies), we can then turn our focus towards increasing PARS's knowledge of threat components. These threat elements/components can be items such as knives, guns, grenades, hatchets, etc. Ideally, a comprehensive training data set of threat components for PARS would be immediately accessible in order to allow it to quickly identify specific weapons and/or threat components, as well as physical objects which could potentially be used as weapons, before determining overall threat levels.

In order to recognize specific threat objects, such as weapons and other dangerous physical objects, an ontological object-recognition classifier can be combined with Argus Prime to improve PARS's threat classification abilities. As a specific example, we can hypothetically add an ontological-based classification (OBC) system, similar to the OBC outlined by Bin Liu, Li Yao, and Dapeng Han in their paper "Harnessing Ontology and Machine Learning for RSO Classification."[14] Ontology-based classifiers exist for a multitude of informational analysis categories, such as natural language processing, written text information retrieval and data mining and medical diagnoses,[15] as well as physical object recognition. OBCs tend to be more effective than classic machine-learning algorithms for object recognition, as ontology classifiers consistently avoid a common machine-learning problem of algorithms overfitting data, which can lead to both inaccurate classifications and cost-function errors.[16]

Additionally, local area information would be necessary for context when it comes to threat components, as good guys carry weapons too. For this, PARS would need to be able to establish familiarity and trust, and I think this could come from the habituation process when acclimating PARS to its operational environment via the LIDA-based reinforced learning approach.

The LIDA-based portion can also implement emotional stressor aspects to be used in conjunction with the classification system already in place to create varying stress levels dependent on the amount of threat components present. These emotional stress levels can achieve the "cognitive tunneling" aspect mentioned previously and prevent less important distractions from influencing PARS during intense situations. For example, if a threat was present and happened to be carrying a hatchet, one AK-47, and two grenades, then a higher threat classification would be applied to that person than to a threatening person who was just carrying one knife. That comparison example should illustrate how the amount of emotional stress in PARS would correlate to the particular threat classification in order to emphasize the severity of a given situation. Lastly, PARS's emotional state would not be influenced solely by threat components present but could also be directly influenced by the number of vulnerable targets present for whom PARS is assigned to protect. For the sake of reassurance—as well as to try and avoid a utilitarian debate similar to the "Trolley Problem"—there probably would be a similar stress level applied toward threats against any amount of vulnerable humans, yet the overall point here is to highlight how a threat analysis process would be undertaken given the increase in vulnerable targets as they relate to PARS's potential "emotional state."

### *GOAL 2: OUTCOME PREDICTABILITY*

The second goal is for PARS to understand its surroundings by analyzing the interactions of objects within those surroundings in complex, nonlinear ways in order to make approximate predictions of what happens next.[17] For effective distraction and prioritization, PARS needs to not only understand the elements that make up threat classifications in goal 1, but it is *imperative* that PARS be able to understand the *probability* of specific outcomes based on those threats. The IPE-modeled system that Battaglia and his colleagues used to determine outcome predictions regarding physical objects would seem to fit our general requirement, and, as previously outlined, the IPE would serve as an important sub-module within the LIDA cognitive cycle. To more clearly understand the concept of physical scene predictability that I am trying to describe, it actually might help to imagine a physics engine (if unfamiliar with what a physics engine is, then I would suggest doing a quick internet search on the topic and viewing some of the video examples that are widely available). Similarly to how a physics engine is able to predict and display simulated physical reactions, the goal for PARS is to be able to accomplish a similar task, but with the purpose of allowing those predictions to influence PARS's priority assessments.

Since approximate probabilistic simulation plays a key role in the human capacity for scene understanding, it is critical that PARS also be able to predict how objects would fall, react when struck by another specific object, resist the force or weight of another object, etc.

Necessary additions outside of just physical scene understanding would also be required for the specific purpose of PARS. These additions would consist of how the specific threat components/weapons a person is carrying operate, as well as what are the threat components' maximum effective range, how many potential targets are vulnerable for attack, etc. Additionally, PARS would need to identify any obstacles that may exist between combatants and targets. Given the success of physics engines like the IPE model outlined by the research team at Massachusetts Institute of Technology, it seems reasonable that a similar framework can be adopted for the purposes of PARS.

### *GOAL 3: PRIORITIZE AND ACT*

Now that PARS is able to (1) notice an object/person/action that is out of place/norm within its operational environment, (2) identify and classify the potential threat level of the element in question, (3) experience an emotional response that emphasizes the severity of the situation and prevents less important distractions from interfering, and (4) make a reliable prediction of what the next event is going to be, PARS should be able to move into the final phase of prioritizing the most important goal within its environment and determine what its next action is going to be.

The LIDA's design is that after observing, identifying, and broadcasting important information across all sub-process networks, the workspace in the cognitive cycle sets out to recruit additional resources to respond to the broadcasts. From there, the cycle moves to goal context hierarchy. This is where the recruited schemes—including emotions—increase their activation and determine an appropriate action. Having given PARS the seemingly innate ability to quickly identify threat components and to predict the most likely physical outcomes, the emotional elements of the LIDA design should begin to influence priorities and action selections based off of those emotional responses. Remember, the emotional attachments should be the product of the procedural learning and familiarization phase of PARS's development. Also, when we hear the words "emotional attachment" we tend to think of a subjective experience that produces something similar to, say, affection, which is misleading in this sense. I only mean "emotional attachment" as an item which would create *any* emotional response within PARS. For example, you may have zero affection for your office computer, but if somebody threw it out of a window, you would most likely have an emotional response to the loss of many important documents contained in that computer. In that example, you might see how your emotional response could be similar to PARS in that in it is most likely the result of an evaluation of a perceived event, and how that event affects you, and your ability to function. Similarly, PARS would develop attachments to people or objects which it is tasked with protecting, and, again, any threat directed at either increases PARS's attention level and inspires PARS to adjust its goals.

## CRITICISM

After hearing this proposal, some people might naturally arrive at the question, "Why not just use LIDA by itself?" I do believe the LIDA framework to be the most useful for our purposes, and after doing research on this topic, I do favor the LIDA designers' approach in emphasizing perceptual learning along with episodic and procedural learning for building emotional attachments. However, for the sake of either immediate practicality, or a failsafe device, or as simply a reassurance provider for a robot functioning in a highly dangerous environment, I do feel that certain innate-like features should be present within the LIDA process.

Outside of just the perceptual, episodic, and procedural learning/memory design of the LIDA, PARS will always retain critical information for quick retrieval, regardless of how closely familiar PARS is with its operational environment. Rather than strict reliance on the processor recruitment design of the LIDA, the goal is for PARS to be able to skip the recruitment process of the most critically important features that pertain to PARS's overall purpose of function (recognizing and reacting to potential threats), thus optimizing response times. Recency/frequency-based memory systems would naturally seem to lag during the processes of problem-solving whenever they encounter elements of a situation that may not be familiar to them, such as unfamiliar weapons or potential threat components. I believe PARS's design can overcome that limitation, as retrieval of that type of specific information would be automatic and threat analysis would continuously occur mandatorily at approximately once every 400 milliseconds.

I also believe this approach has the potential to assist the challenges of trying to get autonomous systems to simultaneously retain focus on an assigned task-oriented goal, while also processing outside world information in a manner which mimics the seemingly innate and subconscious features of human cognition.

Additional criticism may also focus on the current abilities (or inabilities) of technology to achieve the goals I have laid out. Based on personal communication with Troy Kelley, "current robot technology is not capable of identifying things like knives and guns." Outside of object-recognition issues, I am also not sure if the current technology for "novelty detection" is where it needs to be in order to suit PARS's needs. For the purpose of this essay, I am going to leave those challenging elements in, in the hopes that the technology to produce them is not far off. With object-recognition technology continuing to grow by leaps and bounds through new deep learning architectures—such as convolutional neural networks and recurrent neural networks—I am hopeful that the technology needed to address those issues will be available in the not-too-distant future. Additionally, I believe that a more fundamental (or even seemingly natural) approach to object recognition would be better served by heavily focusing on the spatiotemporal aspects of machine learning in the early developmental stage of PARS. Again, just like with human infants, spatiotemporal analysis and anomaly detection is effectively learned and retained, and then is followed by a growth toward feature detection based on those spatiotemporal fundamentals. Therefore, it is not hard to imagine that type of development as being key for quickly advancing object recognition and novelty detection for all autonomous systems.

Lastly, as deep learning mechanisms like convolutional neural networks (CNNs) become loaded with ever increasing amounts of labeled imagery, I am hopeful that weapon types and other potentially hazardous devices will be more easily identifiable and swiftly produce significant advancements in object recognition with regards to machine vision and machine learning.

## SUMMARY

In conclusion, given the necessity of abilities such as distraction and goal prioritization in robots we plan on entrusting with autonomy, certain frameworks are needed to produce those abilities. Given also that the overall intent for PARS was to operate in an environment that heavily

relied on those abilities, it seemed best to ensure that all of the necessary sub-system processors were on hand to produce and reinforce the most critical components of PARS's operations. I feel that the Argus Prime and IPE models serve to do just that, by processing information in a manner similar to innate-like human abilities, while working in conjunction with the current LIDA model to recruit additional and necessary operational processors.

I have not intended that the model presented in this essay be seen as the most ideal format possible for achieving those abilities, but only to show how elements of certain pre-existing models can be used, and perhaps be combined, to provide a more optimal format.

**ACKNOWLEDGMENTS**

This research was supported by a U.S. Army Research Laboratory (ARL) grant to the Philosophy Department at the University of Illinois Springfield (UIS) for research regarding the philosophy of visual processing in object recognition and segmentation. (W911NF-17-2-0218).

I would like to gratefully acknowledge Piotr Boltuc and Troy Kelley for providing continued guidance, expert feedback, and sincere encouragement throughout the entire process of writing this paper. I would also like to thank Brandon Evans for patiently reviewing multiple drafts of this paper.

**NOTES**

1. Kelley and Veksler, "Sleep, Boredom, and Distraction: What Are the Computational Benefits for Cognition?"
2. Oxford Reference, 2018.
3. Franklin et al., "LIDA: A Computational Model of Global Workspace Theory and Developmental Learning."
4. Schoelles, Neth, Meyers, and Grey, "Steps Towards Integrated Models of Cognitive Systems: A Level-of-Analysis Approach to Comparing Human Performance to Model Predictions in a Complex Task Environment."
5. Battaglia, Hamrick, and Tenenbaum, "Simulation as an Engine of Physical Scene Understanding."
6. Baillargeon, "The Acquisition of Physical Knowledge in Infancy: A Summary in Eight Lessons."
7. Scholl, "Objects and Attention: The State of the Art," 36ff.
8. Franklin et al., "LIDA: A Computational Model of Global Workspace Theory and Developmental Learning."
9. Battaglia et al., "Simulation as an Engine of Physical Scene Understanding."
10. Franklin et al., "LIDA: A Computational Model of Global Workspace Theory and Developmental Learning."
11. Madl, Baars, and Franklin, "The Timing of the Cognitive Cycle." Troy Kelley has brought it to my attention that the timing of the human cognitive cycle is around 1 cycle per every 50ms. However, the only research available regarding the timing of the LIDA cognitive cycle shows that its cognitive cycle clocks in at once every 380ms. Given the addition of two new processors for the PARS design, I estimated that an additional 20ms would need to be added to the LIDA cycle.
12. Byrne and Anderson, "Serial Modules in Parallel: The Psychological Refractory Period and Perfect Time-Sharing."
13. Franklin et al., "LIDA: A Computational Model of Global Workspace Theory and Developmental Learning."
14. Liu et al., Harnessing Ontology and Machine Learning for RSO Classification."
15. Khan et al., "A Review of Machine Learning Algorithms for Text-Documents Classification."
16. Liu et al., Harnessing Ontology and Machine Learning for RSO Classification."
17. Battaglia et al., "Simulation as an Engine of Physical Scene Understanding."

**REFERENCES**

Anderson, J., and Schooler, L. "Reflections of the Environment in Memory." *Psychological Science* 2, no. 6 (1991): 396–408.

Anderson, J., M. Matessa, and C. Lebiere. "ACT-R: A Theory of Higher Level Cognition and Its Relation to Visual Attention." *Human-Computer Interaction* 12 (1997): 439–62.

Baillargeon, R. "The Acquisition of Physical Knowledge in Infancy: A Summary in Eight Lessons." In *Blackwell Handbook of Childhood Cognitive Development*, ed. U. Goswami. Oxford: Blackwell, 2002.

Battaglia, P., J. Hamrick, and J. Tenenbaum. "Simulation as an Engine of Physical Scene Understanding." *PNAS* 110, no. 45 (2013): 18327–32. http://www.pnas.org/content/110/45/18327.full.pdf.

Byrne, M., and J. Anderson. "Serial Modules in Parallel: The Psychological Refractory Period and Perfect Time-Sharing." *Psychological Review* 108, no. 4 (2001): 847–69. doi:10.1037/0033-295x.108.4.847.

Cavanna, A., and A. Nani. *Consciousness Theories in Neuroscience and Philosophy of Mind*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

Franklin, S., U. Ramamurthy, S. D'Mello, L. McCauley, A. Negatu, R. Silva L., and V. Datla. "LIDA: A Computational Model of Global Workspace Theory and Developmental Learning." 1997. http://ccrg.cs.memphis.edu/assets/papers/LIDA%20paper%20Fall%20AI%20Symposium%20Final.pdf.

Goswami, U. C., and R. Baillargeon. "Chapter 3: The Acquisition of Physical Knowledge in Infancy: A Summary in Eight Lessons." In *Blackwell Handbook of Childhood Cognitive Development*. Malden, MA: Blackwell, 2003.

Khan, A., B. Baharum, L. Lee, and K. Khan. "A Review of Machine Learning Algorithms for Text-Documents Classification." *Journal of Advances in Information Technology* 1, no. 1 (2010): 4–20. http://www.jait.us/uploadfile/2014/1223/20141223050800532.pdf.

Kelley, T., and V. Veksler. "Sleep, Boredom, and Distraction: What Are the Computational Benefits for Cognition?" *APA Newsletter on Philosophy and Computers* 15, no. 1 (Fall 2015): 3–7. https://c.ymcdn.com/sites/www.apaonline.org/resource/collection/EADE8D52-8D02-4136-9A2A-729368501E43/ComputersV15n1.pdf.

*LIDA Diagram*. (n.d.). https://www.researchgate.net/figure/227624931_fig1_Figure-1-LIDA-cognitive-cycle-diagram.

Liu, B., L. Yao, and D. Han. "Harnessing Ontology and Machine Learning for RSO Classification." *SpringerPlus* 5, no. 1 (2016): 1655. https://doi.org/10.1186/s40064-016-3258-2.

Madl, T., B. Baars, and S. Franklin. "The Timing of the Cognitive Cycle." *PLoS ONE* (2011). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3081809/.

*Oxford Reference*. (2018). http://aut.ac.nz.libguides.com/APA6th/referencelist.

Schoelles, M., and W. Gray. "Argus Prime: Modeling Emergent Microstrategies in a Complex Simulated Task Environment." *Proceedings of the Third International Conference on Cognitive Modeling* (2000): 260–70. http://act-r.psy.cmu.edu/?post_type=publications&p=13921.

Schoelles, M., H. Neth, C. Myers, and W. Gray. (2006) "Steps Towards Integrated Models of Cognitive Systems: A Level-of-Analysis Approach to Comparing Human Performance to Model Predictions in a Complex Task Environment." http://homepages.rpi.edu/~grayw/pubs/papers/2006/07jul-CogSci06/DMAP/SNMG06_CogSci.pdf.

Scholl, Brian J. "Objects and Attention: The State of the Art." *Cognition* 80, no. 1-2 (2001): 1–46. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.547.4788&rep=rep1&type=pdf.

Shah, J. Y., R. Friedman, and A. W. Kruglanski. "Forgetting All Else: On the Antecedents and Consequences of Goal Shielding." *Journal of Personality and Social Psychology* 83, no. 6 (2002): 1261–80. doi:10.1037/0022-3514.83.6.1261.

Tongphu, S., B. Suntisrivaraporn, B. Uyyanonvara, and M. Dailey. "Ontology-Based Object Recognition of Car Sides." Paper presented at the 9th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, Phetchaburi, Thailand, 2012. https://doi.org/10.1109/ECTICon.2012.6254268.

# *Using Quantum Erasers to Test Animal/ Robot Consciousness*

Sky Darmos
HONG KONG POLYTECHNIC UNIVERSITY (POLYU)

## INTRODUCTION

Heisenberg's uncertainty principle, which states that one cannot both know the position and impulse of a particle at once, is not only a restriction for our ability to gain knowledge about nature, but leads beyond that to a general "fuzziness" of all physical entities. By simple interpretation, an electron is not just here or there, but at many places at once. This rather bizarre state is called a superposition.

In the orthodox interpretation of quantum mechanics it is then the *measurement* which leads to a random choice between the various classical states in this superposition. Yet, not all agree upon what constitutes a measurement. Some, such as Heisenberg himself, held that a measurement can't be defined without involving conscious observers.[1] Others, such as Bohr, held that the property of being macroscopic is already enough.[2] But both of them put a strong emphasis on excluding the conscious observer from the observed system.[3] However, in 1932 John Von Neumann wrote a formalization of quantum mechanics and stated that the conscious observer is the only reasonable line of separation between the quantum world and the classical macroscopic world.[4] Eugene Wigner argued the same way in 1963,[5] but withdrew his idea a decade later because he thought it might lead to solipsism due to the way other observers lie on the past light cone of a given observer[6]—a problem which actually can be solved using entanglement.[7] The strong form of the orthodox interpretation (also called Copenhagen interpretation), which explicitly states that it is consciousness which causes the reduction/collapse of the wavefunction, is nowadays referred to as the Von Neumann-Wigner interpretation, or simply as "consciousness-cause-collapse" (CCC).

After the '60s a different view started gaining popularity, namely, that there is no such thing as a collapse of the wavefunction and that we ourselves coexist in a superposition of multiple states as well, each state giving rise to a separate consciousness. It would then be the vanishing wavelengths of macroscopic objects which make the macroscopic world appear rather classical (non-quantum). This interpretation is called many minds interpretation or many worlds interpretation and was popularized in different forms, most noticeably by Stephen Hawking. However, it is important to note that Hawking's version of it is fundamentally different, because there the different "worlds" are put onto separate spacetimes without any causal contact.[8]

It is often held that the above described *measurement problem* is only a philosophical problem and that its various proposed solutions are operationally identical. Students of physics are often told not to worry too much about where and by what means the wavefunction collapses, because interference disappears for macroscopic objects, and thereby, arguably, all means to prove the presence of a superposition.

The basic assumption behind this premise is that even if it is indeed the conscious observer who causes the collapse of the wavefunction, he doesn't have any influence on into which state it collapses. Evidence that this assumption isn't necessarily true doesn't get the attention it deserves.[9]

Even if we put aside all evidence for consciousness being able to influence quantum probabilities, there are still plenty of other ways to test whether or not it is consciousness that causes the reduction of the wavefunction (the choice between realities). Evidence for *macroscopic superpositions* not using interference can be found in various other realms, such as quantum cosmology, quantum biology, parapsychology, and even crystallography.[10] However, *in this paper I want to focus on how to easily test if something has consciousness in a laboratory*, without using a Turing test or any other test for cognitive abilities. These tests might work for human consciousness but are highly inconclusive for other animals.

John A. Wheeler was a strong supporter of "consciousness causes collapse" and one of the first to apply this principle to the universe as a whole, saying, "*We are not only participators in creating the here and near, but also the far away and long ago.*"

How did he come to this conclusion? In the '70s and '80s he suggested a number of experiments aiming to test if particles decide to behave like waves or particles, right when they are emitted or sometime later. For example, one could change the experimental constellation with respect to measuring the path information (polarizations at the slits) or the impulse (interference pattern) after the particle has already been emitted. When the experiments were done many years later, it turned out that what particles do before they are measured isn't decided until after they are measured. This led to Wheeler concluding, "*Quantum phenomena are neither waves nor particles but are intrinsically undefined until the moment they are measured. In a sense, the British philosopher Bishop Berkeley was right when he asserted two centuries ago 'to be is to be perceived'.*"

But many others preferred to rather believe that information partially travels to the past than to believe that reality is entirely created by the mind. Therefore, Wheeler brought the experiment to an extreme by suggesting to conduct it on light emitted from remote galaxies. The experiments showed Wheeler to be right again. The universe indeed materializes in a retrospective fashion.[11]

Later in the '90s new experiments were suggested to test other temporal aspects of quantum mechanics. The so-called *quantum eraser experiment* was also about changing one's mind regarding whether to measure position (particle) or impulse (wave), but here the decision was *not delayed but undone* by erasing the path information.

The erasing is usually not done by deleting data in a measurement apparatus, but simply by undoing the polarization of the entangled partner of a given photon. Polarization doesn't require absorbing a particle. It is therefore no measurement, and the result wouldn't really be introducing much more than Wheeler's delayed choice experiment already did, but there is a special case, namely, undoing the polarization of the entangled partner after the examined photon arrived at the screen already. That is indeed possible, which means the screen itself, although being macroscopic, can be in superposition, at least for short periods of time. *This proves that the screen didn't make the wavefunction collapse*. If we can already prove this, then there must be a way of finding out where exactly the wavefunction collapses.

## USING QUANTUM ERASERS TO TEST CONSCIOUSNESS

Polarizers can be used to mark through which of two given slits, A or B, a photon went, while its entangled partner is sent to another detector. The interference pattern disappears in this situation, but it can be restored if the entangled partner passes another polarizer C, which can undo the marking, resulting in the restoring of the interference pattern. This deleting can be done after the photon arrived at the detector screen, but not long after. Arguably, it is the signal's arrival at the consciousness of the observer that sets the time limit for the deleting.

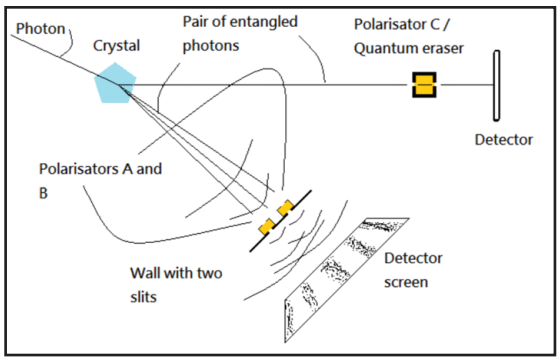


**Figure 1.** Interference pattern reappears when the quantum eraser is used. This happens even if the quantum eraser is further from the crystal than from the screen.

If decoherence theory (or Bohr's scale-dependent version of the Copenhagen interpretation) was right, then the screen should have measured the photon and thereby should have destroyed any chance for interference, simply because it is "macroscopic" (no quantum behavior). Yet that is hard to say because if one doesn't believe in the collapse of the wavefunction (decoherence theory is a no-collapse theory), then interference, and therefore information loss (erasing), may occur at any moment after the measurement.[12,13]

In the Von Neumann-Wigner interpretation it is said that a measurement has to reach a conscious observer in order for the wavefunction to collapse. Yet, if the wavefunction collapsed right in the eye of the observer, there wouldn't be much time for erasing the measurement. Light signals from the measurement arrive almost instantaneously at the eye of the observer (at the speed of light). Thus we can exclude the possibility that the eyeball of the observer causes the collapse of the wavefunction.[14,15]

In my book *Quantum Gravity and the Role of Consciousness in Physics*, I described this experiment and suggested that *one could try to delay the erasing more and more in order to figure out in which moment in time and where in the brain the wavefunction collapses*. It may collapse at a subconscious level already (single projection to the *cerebral cortex* taking less than a half second), or at a conscious level (double projection to the cerebral cortex taking a half second).

It is sometimes suggested that if it is the subconscious which is responsible for the collapse of the wavefunction, then that could explain why we seem to have almost no influence on into which state it collapses.[16]

If erasing the measurement is possible until half a second after the measurement, then consciousness causes the collapse. If this time is slightly shorter, let's say one third of a second, then subconsciousness causes the collapse. We can know this because the temporal aspects of consciousness have been studied quite excessively by the neuroscientist Benjamin Libet.[17]

If we now replace the human by a robot, we would have to place all humans very far away in order to avoid having them collapse the wavefunction. Yet, as soon as the measurement reaches the macrocosm, changes in all fields reach the human with light speed. *And for the wavefunction to collapse, no real knowledge of quantum states needs to be present in the consciousness of an observer. All that is needed is different quantum states to lead to distinguishable states of the mind.*

Another technicality is that although the wavefunctions of macroscopic objects around us collapse every fortieth of a second (the frequency of our brain in the perception realm), the single photons and subsequent brain signals remain in superposition for almost half a second.

When looking at mind over matter interactions, which are mostly about influencing macroscopic systems, the fortieth second is crucial, whereas for quantum erasers, which are about single photons, it is the half second which is crucial.

After testing humans, one can go on and test animals with different brain structure. In some animals the subconscious/conscious level could be reached earlier or later, and that should affect the time limit for the quantum eraser.

Of course, when there is a way to check experimentally if something has consciousness, one can do that for all kinds of things, even robots, cameras, stones, and so forth. It is my belief that something totally algorithmic can't be conscious, simply because such a consciousness wouldn't affect the system's behavior. Only a system which is *quantum random* can have a consciousness that actually affects the system.

Obviously, opinions deviate strongly here, but the good thing is that we don't need to solely rely on beliefs or formal arguments anymore; we can actually go on and experimentally test it.

What we can do is this: Assume that a robot would become aware of things very fast, much faster than the half second it takes for humans. One can then go on and test that by putting the robot in front of the experimental device together with a human. If the robot makes quantum erasing impossible already before the signals reach human consciousness, then the robot is conscious.

Of course, this doesn't account for the possibility that robot consciousness, if existent, is slower than human consciousness (humans experience everything a half second delayed in time!).

Some people think that replacing the human observer by a camera and seeing that the wavefunction still collapses already proves Von Neumann wrong.[18] They miss the point that the quantum state reached the macrocosm already when entering the camera. According to the Von Neumann view, the first time the wavefunction collapsed was after the emergence of life, yet that doesn't have any obvious impact on the world. In Everett's many worlds interpretation the wavefunction never collapses, and again there are no obvious implications. That means only if we try to rapidly erase the measurement can we hope to learn something about where the wavefunction collapses.

In decoherence theory, decoherence replaces the wavefunction collapse. In this theory, objects can be treated classically as soon as interference is lost. Calculating when interference is lost is relatively easy: for any macroscopic object, it is "lost" almost instantaneously. Yet this doesn't tell us when a measurement becomes irreversible. The issue of irreversibility is independent from decoherence (losing of interference), and looking at the ontology of decoherence theory, one would have to assume that erasing a measurement should always be possible. Some took this literally, which led to the creation of rather bizarre theories, such as the "Mandela-effect" where the past is not regarded unchangeable anymore and the universe becomes "forgetful."

According to Max Tegmark, decoherence theory may even lead to a bizarre form of solipsism where consciousness "reads" the many worlds always in a sequential order which leads to its succession—its survival. That is expressed in his thought experiment "quantum suicide." Rather surprisingly, Tegmark doesn't use this to make a case against decoherence theory, but rather wants to show how "thrilling" it is.

## SCHRÖDINGER'S CAT IS REAL

For entities that have a consciousness which is faster than human consciousness, one can easily test that by looking at how much the time window for the quantum eraser is shortened. However, accounting for entities with a slower consciousness, we have to try to isolate the whole system from humans and all other potentially conscious animals. This could be done by moving the whole experiment into a Faraday cage and/or placing it deep beneath the surface of earth and far away from human observers. Nothing that happens inside this Faraday cage should be able to influence anything on the outside.

If the experiment is really perfectly isolated, then the erasing of the which-path information could be delayed further and further. All one would have to do is to let the entangled partner photon continue its travel, for example, by letting it travel circularly inside optical fibers. Yet, if the delayed erasing is to be successful, the entangled partner has to finally hit the third polarizer before the Faraday cage is opened.

Considering how far photons travel in a half second (about 150,000 km), some way to store them without measuring them must be found. Photons travel slower inside optical fiber, reducing the distance traveled in a half second to only 104,927 km, but that is still by far too long for a distance to be traveled in a laboratory. One way to slow them down further could be to let them enter some sort of glass fiber loop. Trapping photons inside mirror spheres or mirror cubes, similar to the "light clocks" in Einstein's thought experiments, is probably not feasible. That is mainly because in such mirror cages photons are often reflected frontal (in a 90-degree angle), and that is when the likelihood of a photon to be absorbed by the mirror is highest (the worst choice here being a mirror sphere[19]). Ordinary mirrors reflect only about half of the photons that hit them. Even the best laser mirrors, so called supermirrors,[20] made exclusively for certain frequencies reflect only 99.9999 percent of the light, and with many reflections (inside an *optical cavity* made of such supermirrors) a single photon would certainly be lost in a tiny fraction of a second. That doesn't happen in a glass fiber wire because there reflection angles are always very flat.[21]

It might prove itself to be very difficult to get the photons in and out of the loop, but even more difficult it seems to get them entering the glass fiber wire in the first place, after they are created together with their entangled partners at the crystal. An option could be to make the glass fiber wire wider at the one end which is used as the entry. One could also guide the photons into the wire by using a focusing lens or a series of guiding mirrors. The first glass fiber wire would lead the photons to the fiber loop. At the place of entry into the loop, the first fiber wire has to be almost parallel to the loop. If the photons always travel in the same direction, they won't ever leave the loop in this case. After sufficient delaying time is gained, the photons have to be taken out and be directed to the third polarizer. That could be achieved if the direction of the entrance fiber wire could be switched so that the entrance becomes an exit. This exit could then be made pointing into the direction of the third polarizer.

In some sense, this experiment would be the first real "Schrödinger's cat" experiment, because just like in Erwin Schrödinger's thought experiment, an animal is put inside a box, here a Faraday cage, and it is theorized about if the animal is in superposition (indicating unconsciousness) or in a certain state (indicating consciousness). But here we have an experimental constellation, which allows us

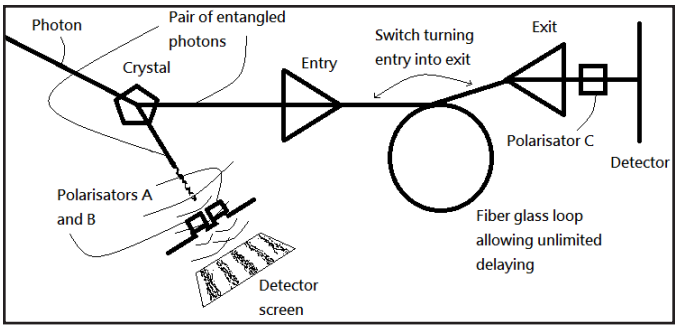


**Figure 2.** Using a fiber glass loop with an entry that can turn into an exit, the erasing of the which-path information can be delayed as much as wished by the experimenter.

to actually check if the animal was in a superposition or not. As for "Schrödinger's cat" in his original thought experiment, one could either just find the cat alive or dead after opening the box. There wasn't any way to tell if the cat had been dead or alive from the beginning, or if it was in a superposition of both states (alive and dead).

## (UNCONCIOUS) ROBOT IN A FARADAY CAGE

For cats, we can be pretty sure that they are conscious, so we can't really make them enter a superposition of being alive and dead at the same time. For robots, that's different: we can be pretty sure that they are unconscious. So if we want to dramatize the experiment, we could have the robot destroying itself when it "sees" an interference pattern.[22] The destruction of the robot (as well as the interference pattern on the screen) could then be erased/undone (!) by the third polarizer. Of course all this has to happen before the Faraday cage is opened. This basically means that the whole past of what happened inside the Faraday cage is decided when it is opened.

However, this is much different from Schrödinger's cat, and maybe much more dramatic. Instead of being in a superposition of destroyed and not destroyed, the robot would "experience" a state of having been definitely destroyed and then a state of never having been destroyed. Of course, that can't be "experienced," and it is just our way of talking about things as if they were real without us looking at them ("looking" here stands for any form of influence to the observer).

A less paradoxical way of talking about this robot is to say that if he destroys himself in the past depends on whether the interference pattern is restored in the future.

## OTHER RESEARCH

### *1. DEAN RADIN AND THE DOUBLE-SLIT-OBSERVER-EFFECT EXPERIMENT.*

In 2016, at the The Science of Consciousness Conference (TSC) in Tucson, Dean Radin gave a lecture which was titled "Experimental Test of the Von Neumann-Wigner Interpretation."[23] Although that was not the name of the associated paper,[24] the experiments he had conducted were basically presented as evidence for consciousness collapsing wavefunctions. Although that has indeed been shown by Radin, the way the experiment was described can be somewhat misleading as to what was really happening. It was a double-slit experiment involving participants "observing" the double slits and thereby altering the *interferometric visibility* of the interference pattern. These human observers were not really watching the double slits with their eyes. They were not staring at the slits to look through which slit the photons passed. If they did so, the photons would go into their eyes and thus we wouldn't have a chance to analyze how the interference pattern was altered. What they did instead is they focused on the slits with their mind. The way Radin puts it, the observers tried to look at the double slits with their "inner eye," in an ESP sort of way. This would be remote viewing; yet one can only remote view things that already exist. A photon that is flying through a double slit does not have a position yet, so the position of the photon is not existing information at that stage.

Therefore, in this experiment the wavefunction is not collapsing any time earlier than usual. It doesn't collapse at the double slit, not even for some of the photons. The wavefunction still collapses only when the photons are registered at the screen and the picture of the screen arrived at the conscious part of the observer's brain.

This experiment is, in its essence, not different from any other micro-PK experiment. Any form of psychokinesis (PK) is proof that something is in superposition, that the wavefunction hasn't collapsed. If somebody can perform PK on, let's say, a cup, it means that the whole cup is in superposition (for a 40th second). Yet if the target object is a single quantum event we speak about micro-PK and all that we can be sure to have been in superposition is the associated quantum particle. However, the observer having an effect on it makes it at least plausible that its quantum state did collapse somewhere in the brain of the observer. In this sense, all nonlocal perturbation experiments can be seen as evidence for consciousness based interpretations of quantum mechanics. Yet, having to deal with so many different interpretations with several of them being related to consciousness, it is obviously not enough to demonstrate the observer effect in order to prove that the orthodox interpretation is the only option.

For some reason, the psi-effect Radin found at the double slits was much stronger than what he and others usually find using other setups such as random number generators (RNG). His result had sigma-5 significance. Maybe the more interesting setup is the main reason for this.

In parapsychology the physical worldview a researcher subscribes to can have a significant impact on how data is interpreted. If someone, in spite of quantum mechanics, believes reality to be based on a time-symmetric space time block universe, for example, he is likely to interpret nonlocal perturbation as precognition.

While I believe the observers were conducting usual micro-PK on the photons, Dean Radin believes the photons were "measured" by remote viewing and the interference pattern was thereby altered. Without going beyond the conventional quantum theory that is afflicted in ambiguity, it will be hard to convince Radin that it was actually micro-

PK and that he should have asked his participants not to mentally "look," but to "wish." A similar debate I have with him about his precognition experiments which I interpret as to represent cases of micro-PK as well (the future picture is selected by a RNG).

He showed that people can react to quantum randomly selected pictures in advance.[25] For me, this is a form of PK. For him, it is precognition. From a general relativity perspective, his opinion makes more sense. From a quantum perspective, PK is the more plausible explanation.

The same also works backwards in time: various researchers have shown that when one uses a computer to record random bits produced by a RNG which are left unobserved for hours, days, and in some cases even for half a year, one still can go and influence the outcome. Looking at this from a space-time perspective, one might suggest that the record in the past was influenced by the observation in the future—an example for retrocausality. And indeed, both Dean Radin and Stephan A. Schwartz argue that way.[26] However, from a quantum perspective, it is more plausible to assume that the record was in superposition all the time before it was played.

An argument against this view by Schwartz is that the success rates are somewhat higher for these retrospective experiments than for ordinary RNG experiments.

Summarizing, we can say that Dean Radin's double-slit-observer-effect experiment can't determine when and where the wavefunction collapses. It is a regular double-slit experiment and that is a thing a regular double-slit experiment just can't do.

Therefore, it is not a test of the Von Neumann-Wigner interpretation to any extent beyond the usual micro-PK experiments.

All we can infer from it is that the observers influenced the outcome. When this influence manifested, we can't know from it. For instance, it doesn't disprove Roger Penrose's gravity-induced wavefunction collapse (OR). What Roger Penrose believes is that it is gravity that induces the collapse, but that it somehow gives rise to consciousness. Others, like Max Tegmark, believe that consciousness chooses its path through an Omnium-like universe of all possible states—an example of this idea is the aforementioned "quantum suicide" thought experiment. These are all examples of theories that don't link the wavefunction collapse to consciousness but that still hold that consciousness has influence over it.

So when testing interpretations of quantum mechanics there are two aspects to consider:

1) Does the observer have an influence on quantum states?

2) When and where does the wavefunction collapse?

Dean Radin's fifty years of research answers (1) with a definite yes, but for answering (2) we need to do the quantum delayed eraser experiment I described here. Fortunately, Radin has just recently expressed interest in conducting the quantum delayed eraser experiment presented here in his lab in the near future.[27]

### *2. LUDOVIC KRUNDEL: DELAYED-CHOICE DOUBLE-SLIT EXPERIMENT OBSERVED BY A ROBOT*

Beginning in 2013, Ludovic Krundel had been promoting an experiment where a robot is looking at a double slit set up with humans staying as far away as possible. He suggested that if the robot is unconscious, then checking through which slit the photons goes shouldn't destroy the interference pattern.

There are several problems with this: firstly, an unconscious robot isn't any different from a normal measurement device, and our experience with measurements is that we can never both obtain the path information and the impulse information (interference).

Secondly, any measurement by the robot would bring the quantum states into the macrocosm and from there it is just a matter of time until the observer's state is influenced.

The way he described it, it was a delayed-choice experiment. Presumably that was influenced by the pre-Wheeler notion of a particle deciding to travel as a wave or a particle before taking off. While accepting the reality of delayed choices, one might think that they cannot happen when the measurement is done by an unconscious robot. It is not too obvious that even when using the Von Neumann criteria of measurement (consciousness-induced collapse of the wavefunction), a measurement doesn't have to be directly displayed to a human in order to count as such. Even in the physicist community, people still sometimes misunderstand the Von Neumann interpretation in this essential way.[28] This is, on the one hand, because pondering about the interpretation problem isn't encouraged much in general, and on the other hand, because Von Neumann himself did not spend much time formulating his interpretation in detail. A clarification that different quantum states only need to lead to different brain states in order to count as measured, without the requirement of any concrete knowledge of these states, would have been very useful. It is this lack of clarity that led to a lot of confusion on if and how to apply quantum mechanics to the macroscopic world.

### RESUME

Why hasn't this experiment been proposed before? One reason is that delaying the erasing for more than just tiny fractions of a second is rather difficult (photons are just too fast). The other reason is that very few physicists are proponents of the Von Neumann-Wigner interpretation and even fewer are familiar enough with concepts in neurobiology in order to link them to things in physics.

And, finally, there is the general misconception that choosing different interpretations doesn't influence predictions on experimental results. We can categorize interpretations of quantum mechanics into scale-

dependent and consciousness-dependent approaches. Most interpretations exist in both variations. We therefore shouldn't really care if there is a wavefunction collapse or a splitting of worlds, because operationally they are the same. All that operationally matters is where the cut is to be placed: Is it scale dependent or consciousness dependent?

It is my opinion that the present results of quantum eraser experiments already prove that scale-dependent approaches can't be right. Some, such as Penrose's gravity-induced wavefunction-collapse theory, might be fine with a detector screen being in superposition for short periods of time. Further delaying the erasing will, however, make it increasingly difficult for any scale-dependent theory to survive.

In my opinion, the interpretation and ontology of a theory is just as important as its mathematical structure. Without a proper interpretation it is not possible to correctly apply the mathematical formalism in all situations. That is just as true for relativity theory. Only by correctly interpreting both theories can a unification be conceived.

In some sense I hold that *pure interpretations* don't exist and that philosophy, correctly done, always leads to hard science.

Note: This is not only an experiment, but can also be turned into a device/product for testing consciousness. The applications would be broad. It could, for example, measure when consciousness is delayed because of drug use.

One who would be perfect for conducting the experiment is the Austrian quantum experimentalist Anton Zeilinger. That is because he is most skilled and renowned in working with interferometers. He could also be good for giving advice on how to conduct the experiment.

**ACKNOWLEDGEMENTS**

Special thanks goes to Professor Gino Yu, who invited me to the CSTS conference in Shanghai (Mai; 2017); Professor Piotr Boltuc, whom I met there, and Dr. Ludovic Krundel, who mentioned my book in connection with testing consciousness in his speech,[29] evoking P. Boltuc's interest and leading up to the creation of this paper.

**NOTES**

1. Werner Heisenberg, *Physics and Philosophy* (George Allen and Unwin, 1958), Chapters 2 (History), 3 (Copenhagen interpretation), and 5 (HPS). Heisenberg says the outcome of the measurement is decided at the measurement apparatus, but the wavefunction doesn't change before the registration in the consciousness of the observer. Although, according to Heisenberg, it is the measurement apparatus where the measurement outcome is decided, the apparatus obtains this power only by being connected to a conscious observer.
2. Niels Bohr, "Unity of Knowledge," in *Atomic Physics and Human Knowledge* (New York, 1958), 73. Niels Bohr never really analyzed the measurement problem. The only hint he gave is that what happens in a measurement apparatus is irreversible, and that is what could constitute a measurement. He insisted that macroscopic objects have to be treated classically but didn't elaborate on why one then can't use macroscopic measurement devises to violate Heisenberg's uncertainty principle. In fact, he had to treat measurement devices as quantum objects before in order to refute some of Einstein's objections and thought experiments in the Bohr-Einstein debate (double-slit experiment with suspended slits measuring tiny displacements in the slit position).
3. This can be said with more certainty for Heisenberg than for Bohr. Although the term "Copenhagen interpretation" is meant to represent the views of both men, it was Heisenberg who formulated the interpretation in a rather unambiguous way and who gave it its name (in 1958). While Bohr often stressed that quantum mechanics allows us only to talk about the outcome of experiments, it was Heisenberg who explicitly stated that observers can't be part of the measured system (see note 1).
4. John von Neumann, *Mathematical Foundations of Quantum Mechanics*, 1932, trans. R. T. Beyer (Princeton University Press, 1996 edition: ISBN 0-691-02893-1).
5. Eugene Wigner and Henry Margenau, "Remarks on the Mind-Body Question," *Symmetries and Reflections*, Scientific Essays, *American Journal of Physics* 35, no. 12 (1967): 1169–70. doi:10.1119/1.1973829.
6. Michael Esfeld, "Essay Review: Wigner's View of Physical Reality," in *Studies in History and Philosophy of Modern Physics*, 30B (Elsevier Science, Ltd., 1999), 145–54.
7. Sky Darmos, "*Quantum Gravity and the Role of Consciousness in Physics*," CreateSpace Independent Publishing Platform, 2014.
8. In this scheme probabilities are re-interpreted as a statistical probability to be in one or the other among many universes.
9. Dean I. Radin, *The Conscious Universe: The Scientific Truth of Psychic Phenomena* (New York: HarperOne, 2009).
10. All this evidence is described in detail in my book, *Quantum Gravity and the Role of Consciousness in Physics*, available both on www.amazon.com and www.academia.edu.
11. Retrospective here doesn't mean that something travels into the past, but that the past is created at the moment of measurement.
12. Though they would claim that information is not something that must be accessible to individuals, but it can be something like the wavefunction of the universe, which is thought of to be out there without being accessible to any particular observer. In this line of thinking, no information is really lost.
13. Decoherence theory can lead to issues with information conservation: If interference is always allowed, then it will happen even with vanishing wavelengths. Within a universe that never experienced a collapse of the wavefunction, quantum probabilities might get lost totally. If the universe is in all possible states right now, then those states should, arguably, all have the same likelihood. In such a world, there would be no reason for an observer to experience a certain succession of states more likely than another.
14. Von Neumann's original paper discussed the question at which place in the brain of the observer the wavefunction might be collapsing.
15. Unless the extra distance travelled by photon is not much longer than the distance of the observer to the measurement device for photon.
16. Lothar Arendes, *Gibt die Physik Wissen über die Natur?: Das Realismusproblem in der Quantenmechanik* (Würzburg, Germany: Königshausen & Neumann, 1992).
17. Benjamin Libet, *Mind Time: The Temporal Factor in Consciousness, Perspectives in Cognitive Neuroscience* (Harvard University Press, 2004). ISBN 0-674-01320-4.
18. Paris Weir, personal correspondence, 2017.
19. Video on the behavior of light in a spherical mirror: https://www.youtube.com/watch?v=zRP82omMX0g.
20. Entry on supermirrors in an encyclopedia of optics: https://www.rp-photonics.com/supermirrors.html.
21. A helpful discussion on trapping photons between mirrors can be found here: https://www.physicsforums.com/threads/light-in-a-mirrored-sphere.90267/.
22. Of course, an interference pattern involves many particles. If only one particle pair is used, then there would be no real pattern,

but still particle A wouldn't arrive at the two possible positions corresponding to straight paths through the slits. That indicates that it interfered with itself. It doesn't really make a difference for the experiment if it is just one pair or many in a row. The erasing works in both cases.

23. TIC 2016 TUCSON, page 194. A video of the lecture can be found here: https:www.youtube.com/watch?v=uSWY6WhHl_M.
24. D. Radin, L. Michel, and A. Delorme, "Psychophysical Modulation of Fringe Visibility in a Distant Double-Slit Optical System," *Physics Essays* 29, no. 1 (2016): 14–22.
25. Dean Radin, *Time-Reversed Human Experience: Experimental Evidence and Implications* (Los Altos, CA: Boundary Institute, 2000).
26. Stephan A. Schwartz, personal correspondence, 2017.
27. Dean Radin, personal correspondence, 2018.
28. Paris Weir, personal correspondence, 2017.
29. Actually, Ludovic Krundel mentioned the possibility of testing consciousness with quantum experiments in connection to my book in all of his speeches since the beginning of 2016. That speech in May 2017 just happened to be the first one I saw from him.

# *The Explanation of Consciousness with Implications to AI*

Pentti O. A. Haikonen
UNIVERSITY OF ILLINOIS AT SPRINGFIELD

In my recent Finnish language book *Tietoisuus, tekoäly ja robotit* (*Consciousness, AI and Robots*),[1] I present a new explanation for phenomenal consciousness. This explanation rejects materialism, dualism, immaterialism, emergentism, and panpsychism. What is left should be self-evident. Here I provide a summary of that argument.

## 1. INTRODUCTION

The brain operates with physical processes that are observable by physical instruments. However, this is not our conscious experience. Instead of percepts of physical processes and neural activity patterns, our contents of consciousness consist of apparently immaterial, phenomenal, qualitative experiences. So far there has not been any good explanation of how the phenomenal experience is generated by the physical processes of the brain.

The problem of consciousness is further complicated by the detection problem; the fact that the actual phenomenal inner experience cannot be detected as such by physical means from outside; it is strictly personal and subjective. So far, instruments have not been able to capture the feel of the redness of a rose, the feel of pain and pleasure, etc. This fact could be taken to prove that firstly, there must be something unique going on, and secondly, the inner experience must be of immaterial nature since it cannot be detected by material means. These conclusions lead to dualistic explanations, where consciousness is seen as a separate immaterial substance or some emergent non-material mental property. These explanations are not satisfactory.

An acceptable explanation of phenomenal consciousness would explain how the inner phenomenal experience arises without resorting to dualism or emergence. Here I give such explanation based on the physical perception processes in the brain.

## 2. PERCEPTION AND QUALIA

All our information about the physical world comes via our senses. The brain operates with neural signals, and consequently it is not able to accept non-neural external stimuli, such as sound, photons, temperature, odor, taste, etc., as direct inputs. Therefore, senses transform externally sensed stimuli into neural signal patterns that convey the sensed information. The resulting signal patterns are not the sensed entity or property itself; instead, they are neural responses that are generated by the sensors' reactions to the sensed stimuli. Consequently, the eventual phenomenal percepts are not the actual properties of the sensed phenomena; instead, they are kinds of "false color" impressions of these. The experienced sweetness of sugar is not a property of sugar; instead, it is the evoked reaction of the system. The experienced redness of a rose is not a property of the rose; instead, it is the evoked reaction of the system to the excitation of the cone cells in the retina by certain photon energies.

The important point here is that we do not experience these reactions as neural activity. Instead, these neural activities appear internally as apparent qualities of the world: sounds, visual forms, colors, odor, taste, pain, pleasure, etc. These sensations are called qualia. More generally, whenever any neural activity manifests itself as a percept, it manifests itself as a quale, not as the actual neural activity.

This leads to the big question: Why and how does some of the neural activity in the brain manifest itself as qualia and not as the actual neural activity as such or not at all? This question is known as "the hard problem of consciousness" as recognized by Chalmers[2] and others, and the solving of this problem would constitute the explanation of phenomenal consciousness. The issues that relate to the contents of consciousness, such as self-consciousness, situational awareness, social consciousness, etc., are consequential and do not have a part in the explanation of the basic phenomenal consciousness.

## 3. ARE QUALIA NON-PHYSICAL?

It is generally understood that at least in principle, all physical processes can be detected and measured by physical instruments via physical interactions between the detector and the detected. Accordingly, various physical brain imaging methods are able to detect neural activity patterns and neural signals in the brain. However, no instrument has ever been able to detect qualia. Pain-carrying neural signals can be detected, but the actual feel of pain remains undetected. The same goes for all qualia. Phenomenal experiences cannot be detected by physical instruments. Surely, this should show that qualia and consciousness are non-physical, immaterial entities, or would it? On the other hand, if it could be shown that qualia were not immaterial, dualistic explanations of consciousness would be unnecessary.

This problem can be solved by the scrutinization of the general process of measuring. Measuring instruments and arrangements detect and measure only the property that they are designed to measure. If you measure a photon as a particle, the photon will appear as a particle. If you measure a photon as a wave, the photon will appear as a wave. However, the particle view and the wave view are only our own models and descriptions of the photon, while the photon as itself is what it is. Measurements do not reveal the actual photon as itself, "das Ding an sich." The same goes for all measurements. The measured object is not revealed as itself; instead, our instruments give some symbolic patterns and values that represent and describe some properties of the measured object. Therefore, the failure to detect and measure qualia is not a unique situation. Instead, it is the direct consequence of the universal limitations of detection and measurement processes. It is not possible to externally access the detected entity as the phenomenal itself, and the only instrument that can detect phenomenal qualia is the experiencing system itself. Consequently, the undetectability of qualia is not an indication of any non-physical nature of the same.

Based on the above, it should be obvious why sensory neural activities appear as qualia instead of appearing as actual neural processes. There is no reason why the neural sensory responses should internally have similar material expression that we get from the outside by our instruments in the first place. In the brain there are no sensors that could detect neural signals as such, and if there were, the neural signals would not be detected as themselves, but as the reactions of the detecting sensors.

Neural sensory responses result from the inspection of the world by senses and consequently, the responses are not about themselves; they are about the sensed stimuli and assume qualities of the stimuli, albeit in a different form, like false color imagery. The mind is not able to access the world as "das Ding an sich" any better than we are with our instruments. Yet, we believe that we perceive the world exactly as it is and our impressions of colors, sounds, smells, etc., are actual world properties. They are not; they are the way in which the neural sensory responses are experienced internally. Technically, this is not much different from the radio, where the radio frequency carrier wave carries the transmitted sound as modulation.

## 4. PERCEPTION, QUALIA, AND CONSCIOUSNESS

The content of consciousness is always about something. It may consist of percepts of the external world and the physical body or thoughts, memories, and feelings, or the combination of these. Introspection shows that superficially the contents of consciousness always appear in terms of sensory percepts, which in turn have the form of qualia.

Inner speech appears as a kind of heard speech, imaginations appear as seen images, imagined actions appear as being virtually executed and perceived by proprioceptors. This kind of effect can be produced by internal feedback loops that return the products of mental processes into virtual percepts.[3,4,5] Without this feedback process, the products of mental processes would not become consciously perceived because in the brain there are no sensors that could sense the neural activity as such. And if there were, it would be no good, as the neural activity as such is not interesting, only the carried information matters. And this can be decoded by returning it into virtual percepts.

The qualia-based percepts generated by sensory perception indicate the instantaneous presence of the corresponding stimuli: seen objects, heard sounds, smell, etc. Without any additional mechanisms, these percepts would disappear without a trace as soon as the stimuli were removed. However, in conscious perception the percepts can be remembered for a while. They can be reported verbally or by other means, and they can evoke various reactions and associations, and this very action separates conscious perception from non-conscious perception. The effect of a conscious percept goes beyond the automatic stimulus-response reaction. The required additional mechanisms are short-term memories and associative long-term memories with the aforesaid feedback configuration. This is an easily implementable technical requirement, and as such does not call for any ontological explanation.

Qualia are self-explanatory; they do not need any interpretation. Red is red, visual patterns are visual patterns, pain hurts directly, a hand position is a hand position, and no names or additional information are required to experience them. Their appearance and feel are their intrinsic meaning. However, additional meanings can be associated with these sensations. These additional associated meanings, such as names and affordances, allow the generation of mental concepts and their mental manipulation. Technically, this calls for associatively cross-connected neural network architectures. These architectures can be created by artificial means.[6]

An important form of the contents of consciousness is the inner speech that uses a natural language. A natural language is a symbolic system with words as symbols. It is known that in closed symbolic systems, such as natural language or mathematics, the meanings of the used symbols cannot be ultimately defined by other symbols within the system. Syntactic operations will not lead to semantics, as pointed out by, e.g., Searle.[7]

A natural language is a method for the description of the external world, and therefore the used words must ultimately refer to external entities and conditions; the meanings of the words must come from outside the symbolic system. However, this outside information cannot be in the form of symbols because these would only enlarge the original symbolic system, and the number of symbols to be interpreted would only increase. Successful grounding of meaning calls for self-explanatory pieces of outside information. It should be evident what the forms of these self-explanatory pieces of information would be; they are qualia.

## 5. THE EXPLANATION OF CONSCIOUSNESS

The author argues that consciousness is not any material substance. Furthermore, the author argues that consciousness is not an immaterial substance either, such as a soul or panpsyche. Obviously, this approach eliminates all dualistic explanations.

It is argued that 1) consciousness is perception with self-explanatory qualia and short-term memory that allows reportability. Without percepts the contents of consciousness is empty; there is no consciousness. 2) Qualia are the way in which the neural sensory responses are experienced by the system itself. Consequently, they are "das Ding an sich" that can externally be observed only as neural activity and not as any phenomenal "feel."

The rejection of dualism: Technically, perception is interaction consisting of the flow of neural sensory responses that associatively evoke other neural activity patterns. Action and interaction are not a material or an immaterial substance any more than the raising of a hand or running. The assumption of otherwise leads to category error and to attempted dualistic explanations that in the end try to explain what is to be explained by the unexplainable.

## 6. IMPLICATIONS TO AI

True general intelligence calls for true understanding. This can only be achieved by the grounding of the meaning of the used symbols to the external world—its entities and conditions. This in turn calls for perception processes. Contemporary computers do have cameras and microphones and possibly other sensors, but they always transform the sensed information into the digital currency of operation, namely, binary numbers. These are symbols without any intrinsic meaning, and the computer manipulates these as any calculator would. The numbers mean nothing to the computer, and the interpretation of meaning remains to the human operator. The grounding of meaning remains missing.

It was argued here earlier that the grounding of meaning calls for external information that is self-explanatory, and this kind of information has the form of qualia. Consequently, eventual machines that understand and operate with external meanings must have perception processes that produce percepts in the form of qualia. These qualia do not have to be similar to human qualia. To have perception process with qualia is to have consciousness; thus, true intelligent machines will have to be conscious.

### NOTES


1. P. O. Haikonen, *Tietoisuus, tekoäly ja robotit* (Helsinki, Finland: Art House, 2017).
2. D. Chalmers, "Facing Up to the Problem of Consciousness," *Journal of Consciousness Studies* 2, no. 3 (1995): 200–19.
3. P. O. Haikonen, *The Cognitive Approach to Conscious Machines* (UK: Imprint Academic, 2003).
4. P. O. Haikonen, *Robot Brains* (UK: Wiley, 2007).
5. P. O. Haikonen, *Consciousness and Robot Sentience* (Singapore: World Scientific, 2012).
6. Ibid.
7. J. R. Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3, no. 3 (1980): 427.


# *Digital Consciousness and Platonic Computation: Unification of Consciousness, Mind, and Matter by Metacomputics*


Simon.X.Duan

**METACOMPUTICS LABS, UK**


## INTRODUCTION

Throughout the history of human civilization, driven by our never-ending curiosity, many ideas have been proposed to explain the world we live in.

Observation of the world gives us conceptual metaphors that are often used to propose theories and models. Light as a wave, light as particles, gas as billiard balls, electric current as flow and the atom as a planetary system are all examples of metaphor-based hypotheses that have been accepted as mainstream scientific theories. Many others, including the plum pudding model of the atom, were discarded when they failed to explain new experimental results.

Since the second half of the twentieth century, inspired by the development of computation and telecommunication technologies, some computer scientists and physicists have proposed new ideas of the world that can be categorized by the terms *digital physics* and *digital philosophy*.

These theories are grounded in one or more of the following hypotheses that the universe

- is essentially informational
- is essentially computable (computational universe theory)
- can be described digitally
- is in essence digital
- is itself a computer (pancomputationalism)
- is the output of a simulated reality exercise

Konrad Zuse (1969), one of the earliest pioneers of modern computer, first suggested the idea that the entire universe is being computed on a computer.

John Wheeler (1990) proposed a famous remark "it-from-bit":

> "It from bit" symbolizes the idea that every item of the physical world has at bottom—a very deep bottom, in most instances—an immaterial source and explanation; that which we call reality arises in the last analysis from the posing of yes–no questions and the registering of equipment-evoked responses; in short, that all things physical are information-theoretic in origin and that this is a participatory universe.

The terms *digital Physics* and *digital Philosophy* were coined by computer scientist Edward Fredkin (1992) who

speculated that it (Fredkin, 2005, p275) "only requires one far-fetched assumption: there is this place, Other, that hosts the engine that 'runs' the physics."

Related ideas include the binary theory of ur-alternatives by Carl Weizsäcker (1980) and ultimate ensemble by Max Tegmark (2007).

Others who have modeled the universe as a giant computer include Stephen Wolfram (2002), Juergen Schmidhuber (1997), Hector Zenil (2012), and Tommaso Bolognesi (2012).

Quantum versions of digital physics have been proposed by Nobel laureate Gerard 't Hooft (1999), Seth Lloyd (2005), David Deutsch (1997), Paola Zizzi (2005), and Brian Whitworth (2010).

Greg Chaitin (2012) suggested that biology is all about digital software. Marcus Hutter (2012) proposed a subjective computable universe model which includes observer localization.

The previous works, however, have not considered how such a giant computer capable of calculating the universe could have come into existence.

This paper proposes a metaphysics framework that provides a foundation to support digital physics and digital philosophy hypotheses.

The metaphysics approach is necessary to establish a Platonic computation system outside the physical universe in order for it to construct and operate the physical universe. This belief is based on the idea, as Albert Einstein said, that "no problem can be solved from the same level of consciousness that created it."

Proposed below is a metaphysics model that uses Platonic objects to describe the creation of the Metacomputation System (MS). This MS consists of three faculties (data, program, and processor) that construct and operate the processed existence.

Through the convergence of computation theories and metaphysics, the proposed model clarifies a range of important concepts and phenomena that cannot be explained by existing accepted theories.

## DESCRIPTION

The *Metacomputation System* (MS) is derived from a metaphysics model based on the following premise:

> There exists *Source Mind*. *Source Mind* is the potential power to conceive, to perceive, and to be self-aware.

Source Mind is one aspect of *Life*. Other imaginable aspects of *Life* such as unconditional love, joy, beauty, and benevolence, as well as its unimaginable aspects, are beyond the scope of this model.

Using the following descriptive terms, we can get a sense of what Source Mind is not:

Timeless, non-spatial, dimensionless, infinite, boundless, non-dual, formless, no-thing, non-changeable, non-destructible, non-comprehensible, non-describable.

The content of Source Mind has a three-tier hierarchy structure constructed with Platonic objects described as follows.

### *UNITY TIER*

The most fundamental creation that Source Mind conceives is *Unity Screen*, represented in Figure 1.

Unity Screen is created so that Source Mind can express itself in form; by projecting itself onto Unity Screen, Source Mind makes itself perceivable.

Unity Screen is of the size of one unit. It contains one pixel of the projected power of Source Mind.

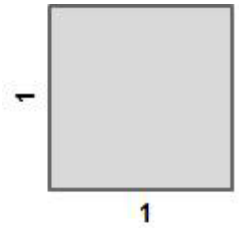


**Figure 1.** Unity Screen that contains one pixel of the projected power of Source Mind.

The nature of existence at unity tier can be described as one, uniform, even, equal, neutral, stable, non-changing, constant, still, singular, total.

### *DUALITY TIER*

At the duality tier, Unity Screen is divided into four cells of equal size as illustrated in Figure 2.

Unity Screen of one pixel is then split up into two symbols: A and B, as illustrated in Figure 3.

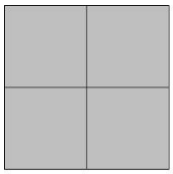

**Figure 2.** Division of Unity Screen into four cells of equal size.

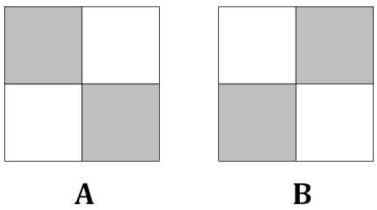


**Figure 3.** Symbols A and B derived from dividing the pixel in Unity Screen. Each symbol contains two pixels and two voids in polar opposites.

Each of these symbols contains two pixels and two voids.

A *void* is a cell within Unity Screen that contains the potential power of Source Mind but is absent of the projected power of Source Mind.

Thus duality is conceived as the polar opposite of the potential and projected power of Source Mind. Void represents potentiality whereas pixel represents actuality.

### *CONCEPTION OF CHANGE*

As Unity Screen (see Figure 1) defines the limited scope of perception of Source Mind, the two separate symbols A and B (Figure 2) can no longer be perceived at the same time. Thus the two symbols are to emerge in Unity Screen in temporal sequence one after the other.

The alternating appearance of symbols A and B can be imagined to be brought about by a looped movement of the inter-connected symbols A and B from right to left as illustrated in Figure 4.

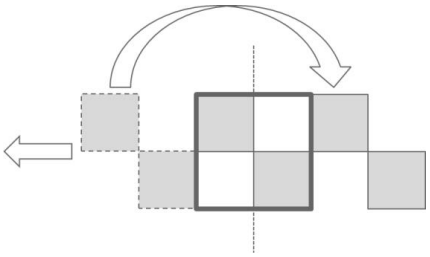

**Figure 4.** Looped movement of the inter-connected symbols A and B across Unity Screen (outlined with thick lines).

From this point of view, when the inter-connected symbols A and B move across Unity Screen, each cell within Unity Screen switches from one state (pixel or void) to the opposite state.

Thus a clock is perceived from the perspective of Unity Screen with its four cells alternating between the two opposite states.

At the first half-clock cycle, symbol A switches to symbol B; at the second half-clock cycle, symbol B switches to symbol A.

The passage of the inter-connected symbols A and B creates temporality. Temporality is measured using *Unit*.

1 Unit = the width of Unity Screen

*Present Moment* (**PM**) is defined as the temporal duration for one switching cycle to complete.

At the duality tier,

**PM** = 1 Unit.

Clock speed = 1 cycle/Unit.

Change, movement, switch, and clock are thus derived at the duality tier and perceived by Source Mind.

The nature of existence at duality tier can be described as follows: changing, moving, dynamic, and rhythmic.

### *TRINITY TIER*

In Figure 2 Unity Screen of one pixel is divided into four pixels in four cells. Each pixel can be further divided into four pixels in four cells.

This sequence of division and resulting duration of PM can be described as follows:

$$\{1, \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{16}, \tfrac{1}{32}, \tfrac{1}{64}, \tfrac{1}{128}, \tfrac{1}{256}, \tfrac{1}{512}, \ldots\} \text{ Unit}$$

Suppose the number of times Unity Screen is divided = N, then,

PM = $2^{-(N-1)}$ Unit

Clock Speed = $2^{(N-1)}$ cycles/Unit.

PM is represented by the shaded cells in the center of Figure 5 where Unity Screen is divided six times.

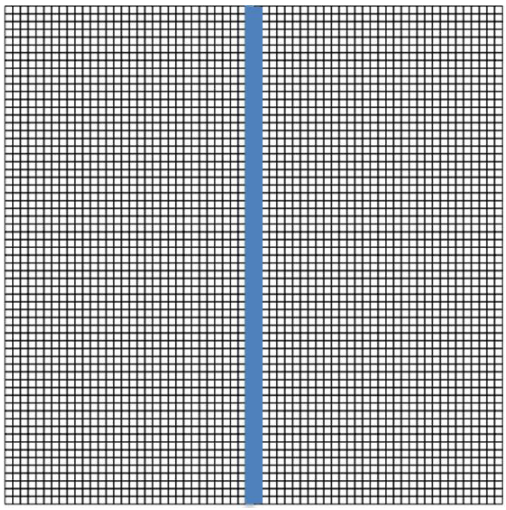

**Figure 5.** MS Grid showing Unity Screen is divided six times. The shaded cells represent PM.

The number of cells produced by each division is as follows:

{4,16,64,256,1024,4096,…}

Suppose the number of times Unity Screen is divided = N, then,

Number of cells = $4^N$,

As each cell can be used to store binary data by assigning a pixel as 1 and a void as 0, thus,

Memory of the grid = $4^N$ bits.

It should be noted that the cells in the PM are operating switches. Thus, in the PM,

Number of operating switches = $2^{N+1}$

### *CONCEPTION OF METACOMPUTATION SYSTEM (MS)*

The availability of sufficient number of switches and memory derived from the grid in Figure 5 (named MS Grid) enables the creation of the *metacomputation system* (MS) that consists of the following three faculties:

- ***Data*** – Specific configurations of pixels (1s) and voids (0s) in binary opposites derivable from the MS Grid.
- ***Program*** – Sequences of codes in binary opposites derivable from the MS Grid that instruct the processor to process data and output results.
- ***Processor*** – Purposefully configured set of pixel/void switches derivable from the PM in the MS Grid that enables arithmetic and logic operations and memory functions. It accepts data, performs instructed computations, and outputs results. A clock is used to regulate the speed of computation.

The MS is a moving grid of cells of pixel/void passing a fixed window of PM. MS contains data, program, and processor. Computation occurs at PM.

The MS is created, sustained, and powered by Source Mind.

## DISCUSSION

### *CONSTRUCTION OF PROCESSED EXISTENCE*

Figure 6 illustrates the proposed mechanism of creation in which the MS is derived from a three-tier hierarchy of Platonic objects conceived by Source Mind.

In Figure 6, each subsequent tier is a derivative of the previous substrate tier. Existence increases its complexity when the derivative tier is conceived.

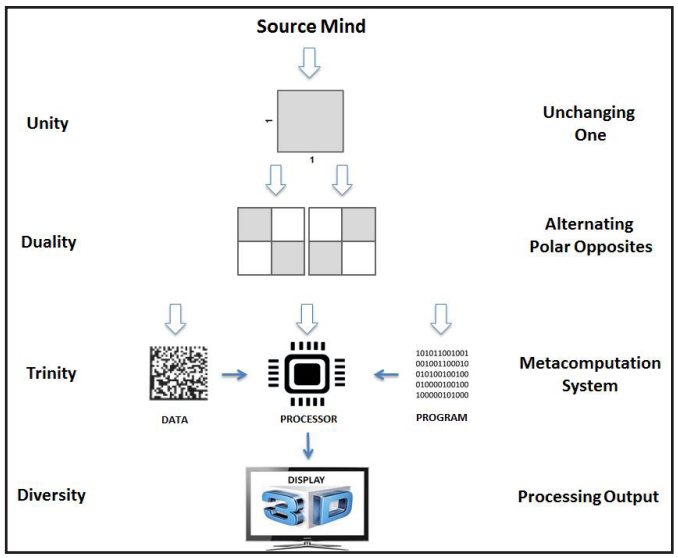


**Figure 6.** Mechanism of creation in which the MS is derived from a three-tier hierarchy construct of Platonic objects conceived by Source Mind. The resulting MS constructs processed existence as its processing output.

The derived MS consists of three faculties: data, program, and processor.

These three faculties interact to construct the processed existence including time, space, and all its content.

This is modeled from our daily observation in this digital age. For example, a DVD disc contains data, but only when it is put into an operating computer and processed with programs can the image and sound then be perceived.

According to this model, all our perceptions and experiences are processing outputs of the MS. This will be discussed in more detail in the following sections.

### *TIME*

Figure 7 is a segment taken from the MS Grid in Figure 5.

As shown in the graph, interconnected symbols A and B (see Figure 3) form a square wave of alternating pixels and voids. The waveform can be likened to the clock signal used in electronic computers.

*Present Moment* is a window from which perpetual progression of the pixel square wave from right to left is perceived. The position of the window is arbitrary and can be fixed anywhere in the MS Grid.

*Future* is represented by the parts of the pixel square wave that are moving towards but have not yet arrived at present moment; *Past* is represented by the parts of the pixel square wave that have moved away from present moment.

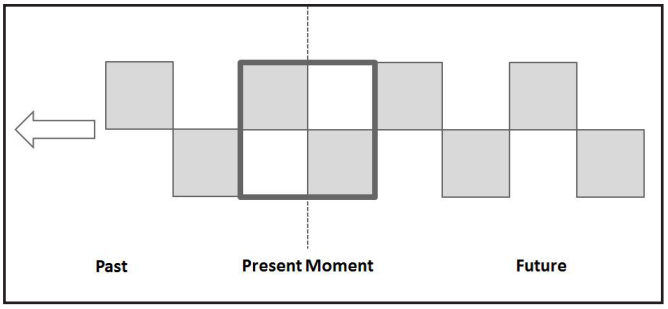


**Figure 7.** Illustration of *Time* as the perpetual progression of the pixel square wave that completes one switching cycle in PM.

Within PM outlined by the thick line in Figure 7, each of the four cells completes a full switching cycle at every $2^{-(N-1)}$ Unit.

PM is the moment when switching, and therefore computation, takes place.

*Time* is thus defined as one-directional perpetual progression of the pixel square wave that completes one switching cycle in PM.

The pixel square wave that defines time in Figure 7 can be expressed as two rows of *time bit strings* of perfect regularity:

……101010101010101010……

……010101010101010101……

Time bit strings can be regarded as a program. Time is perceived when the program is executed.

### *SPACE*

Unity Screen in Figure 1 defines the scope of temporality in horizontal direction. It also defines the scope of dimensionality in vertical direction.

The progression of the pixel square wave in time in horizontal direction at PM is associated with propagation of the pixel square wave in vertical direction. This is illustrated in Figure 8.

Thus the absolute space in vertical direction at PM is filled with alternating pixels and voids.

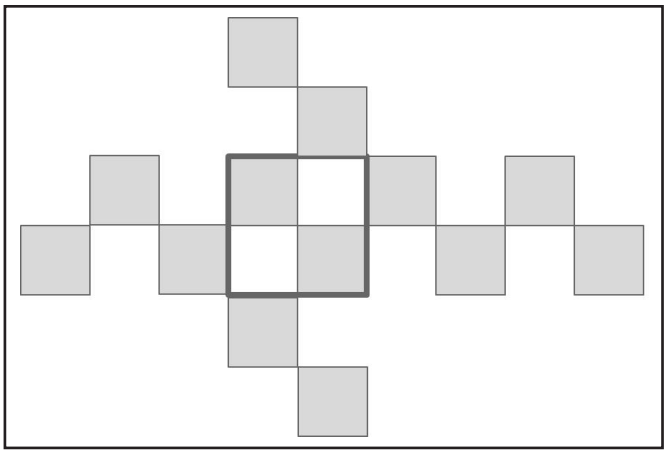

**Figure 8.** Propagation of the pixel square wave in vertical direction in the absolute space is associated with progression of the pixel square wave in time in horizontal direction at PM.

A program can be deployed to create 2D coordinates using time bit string in both an **X** and **Y** axis.

Figure 9 illustrates a section of the 2D space thus constructed.

It can be seen that the 2D space is formed by perfect regular arrangements of alternating pixels and voids.

Figure 9 is the state of the 2D space at a given half cycle moment in time. At the next half cycle moment each pixel and void switches to its opposite.

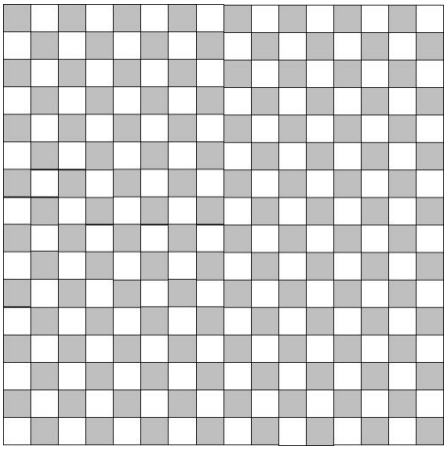

**Figure 9.** 2D space constructed by using time bit string in an X and Y axis. The shaded cells are pixels and light cells voids.

Similarly, a program can be deployed to create 3D coordinates using time bit string with an additional **Z** axis.

With such program, a 3D grid as illustrated in Figure 10 is constructed.

It should be noted that the pixels represented in the 2D space grid in Figure 8 are transformed into voxels charged with the power of Source Mind.

A powered voxel is named a *poxel*.

*Poxel* is the 3D expression of the power of Source Mind in space.

According to the model, space is a 3D grid filled with regularly patterned poxels and voids. Figure 9 is a section of 3D space at a given half cycle moment in time. At the next half-cycle moment, each poxel and void switches to its opposite.

Thus, space is not empty—instead, it is filled with regularly patterned alternating poxels and voids.

As Space is constructed using pixel square wave and time bit string, it can be said that Space is a derivative of Time.

Space also functions as a 3D display. The processing output of the MS is displayed in the 3D space.

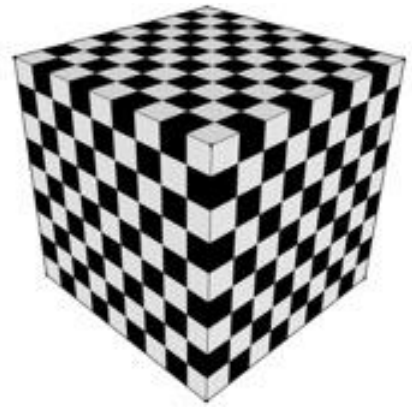

**Figure 10.** 3D space represented as 3D grid. The dark voxels are poxels and the light voxels voids.

For instance, programs can be executed to output into space points, lines, plains, shapes and other forms of abstract objects. These objects are printed in space using poxels.

### *LEVELS OF CREATION AND MULTIVERSE*

In the MS Grid, different **N** values can be used to create multiple MSs. Each MS with a different **N** value operates at a different clock speed according to the formula below:

$$\text{Clock speed} = 2^{(N-1)} \text{ cycles/Unit}$$

It can thus be assumed that many levels of creation are in existence. Our physical universe is one of many parallel universes.

A universe produced by the MS operating with a bigger **N** value is equipped with a more powerful processor and has more memory to accommodate larger quantities of data and programs. It therefore allows richer and more diverse perceptions and experiences.

It should be noted that the position of PM in Figure 5 is arbitrary. It can be positioned anywhere in the grid. Therefore, the entire history of creation at all levels can be computed.

We assume the physical universe is a processing output of the MS operating with **N** value. Levels of creation produced by the MS operating with smaller **N** values are viewed as higher levels of creation.

Ascending the levels of creation implies experiencing the universes produced by the MSs operating with a smaller **N** value.

Figure 11 illustrates a selection of 3 MSs in the multiverse.

At the top level, **N** = 1,

**PM** = 1 Unit; Clock speed = 1 cycles/Unit

At the middle level, **N** = 4,

**PM** = 1/8 Unit; Clock speed = 8 cycles/Unit

At the lower level, **N** = 6,

**PM** = 1/32 Unit; Clock speed = 32 cycles/Unit

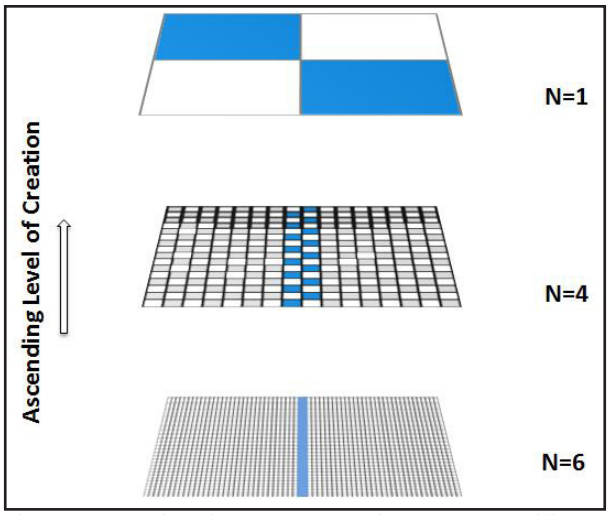


**Figure 11.** Selection of three MSs operating at the three different clock speeds. PM (colored blue) decreases with increasing **N** values.

## *CREATION OF ENTITIES*

*Entity* is a being with both subjective and objective aspects. For instance, a human being is an entity having both a mind (the subjective aspect) and a body (the objective aspect).

The objective aspect of an entity is the processing output of the MS displayed in space as a 3D image named *Entity Image* (EI). EI is determined by a specific dataset, as well as the programs and the processor that are deployed to produce the output.

Poxel is the building block of EI. EIs are created by arranging the poxel in specific configurations and patterns that deviate from the regularity exhibited by space.

In this digital age, perceiving images on screen is part of modern day living. For example, a mobile phone receives digital data in the form of 1s and 0s. They are then processed using programs. The processing output is the image displayed on the screen of the phone.

Likewise, entities can only be perceived as meaningful forms when the dataset of an entity is processed by the programs in the MS.

A given physical entity exists at every other level of creation and is perceived as different EIs at the different levels of creation.

With an increasing **N** value more powerful processors become available. The dataset of an entity as well as programs available increase in size and complexity.

With more complex data and programs that give properties to EIs, such as mass, solidity, transparency, color, texture, richer features of the EI can be perceived.

The physical form displayed at the physical level of creation is a complex EI of a given entity. At higher levels of creation (with a smaller **N** value) simpler non-physical EI is perceived.

Entities can be categorized in different ways, for example:

**By size and composition:**

Universe, galaxy, planets, material object, cell, molecule, DNA, etc.

**By state:**

Solid, liquid, gas, plasma, etc.

**By complexity:**

Human, animal, plant, mineral, air, water, etc.

The subjective aspect of an entity is its mind (see section **Mind**).

## *DILATION OF TIME*

From the definition of Present Moment (PM), it is established that

$$\mathbf{PM} = 2^{-(N-1)} \text{ Unit.}$$

PM decreases with the increase of the **N** value.

Suppose the physical universe is produced by the MS operating with a value $\mathbf{N_P}$, PM in the physical level of creation is of the value $\mathbf{PM_P}$,

We call the level of creation that is **m** level higher than the physical universe level **m**, then

$$\mathbf{N} = \mathbf{N_P} - \mathbf{m},$$

$$\mathbf{PM_m} = 2^{-(Np - m-1)} \text{ Unit,}$$

Thus,

$$\mathbf{PM_m}/\mathbf{PM_P} = 2^{-(Np - m-1)} \text{ Unit}/2^{-(Np -1)} \text{ Unit} = 2^m$$

**PM** at level **m** is $2^m$ times that of the physical level creation.

Suppose **PM** = 1 (Day). Then,

1 (Day) **m** level time = $2^m$ (Day) physical level time

### *LANGUAGE*

Program is identified by giving a name to it. Specific words are intended to name specific programs. The true meaning of a word is the perception experienced from executing the program.

For example, Space is perceived by running program {Space}.

Light is experienced when program {Light} is executed to produce specific poxel waves in space.

Redness is perceived when program {Red} is executed.

{Apple} identifies a program that enables the concept "Apple-ness" to be perceived.

Names of complex programs giving meaning to entities in creation include the following:

- Cosmological objects: galaxy, planet, etc.
- Physical matter: solid, liquid, gas, plasma, etc.
- Biological systems: plant, animal, human, cell, etc.
- Programs are used to define the meanings of abstract concepts.

The meaning of number, for example, {2}, is perceived when a successor program is executed with 1 as the initial state.

{Mass} is a program that defines the inertia of an object to change its state of motion in space.

{Force} is a program that defines the cause for an object to change its state of motion in space.

{Heat} is a program that defines the dynamic property of a system.

{Energy} is a program that defines the capacity of a system to do work.

Other programs include the descriptive terms used in human languages. These programs allow the human mind to experience a wide range of thoughts, emotions, feelings, sensations, actions, and interactions.

The evolution of human civilization is marked by development of programs. The creation of each new word corresponds to the availability of a new program to the society where the word is used.

Programs are stored in the memory of the MS and can be identified and retrieved through the use of language.

### *LIFECYCLE OF ENTITIES*

We have established that the memory of the MS at level **N** = $4^N$.

As a computation system with finite memory, its processing output cannot increase indefinitely. This leads to a logical conclusion that entities have to go through a life cycle and have a limited life span.

All entities run program {life cycle} that progresses them through the stages of inception, expansion, deterioration, and termination in time.

It is assumed that at a given level of creation, an EI has a life span determined by a fixed number of processing cycles (or fixed number of **PM**s) from its inception to termination.

As each level of creation is constructed by computation at different clock speeds, each EI's life span at a different level of creation will be different for a given entity.

For instance, for a given entity, if the life span of its EI at the physical level

$$\mathbf{L_P} = k\ (\mathbf{PM_P})$$

Then the life span of its EI at level **m**

$$\mathbf{L_m} = k\ (\mathbf{PM_m}) = k \times 2^m\ (\mathbf{PM_P})$$

The entity thus experiences $2^m$ times as long a life span with its EI at level **m** compared to its EI at the physical level.

For a given entity, its EI's life span at a different level of creation can be illustrated as a hierarchy shown in the example in Figure 12, where $\mathbf{L_p}$ is the life span of the EI at the physical level, $\mathbf{L_{p-2}}$ is the life span of the EI at 2 levels above the physical level and $\mathbf{L_{p-4}}$ 4 levels above the physical level.

For a given entity, with a descending level of creation (increasing **N** value), multiple EIs with shorter life spans exist consecutively in time.

The life span of its higher EI is the sum of all the life spans of its lower EIs.

Many EIs at a lower level of creation can correspond to one EI at a higher level of creation.

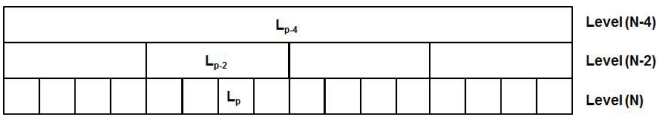


**Figure 12.** Example of the relative life span (L) of a given entity at different levels of creation.

### *MEMORY OF MS*

Theoretically, Planck time is the smallest meaningful unit of time in the physical universe.

If we assume:

Width of the pixel = Planck time,

Time span of perceivable creation

= Size of Unity Screen

= Life span of the physical universe

= (13.8 + 5) billion years,

Then,

$\mathbf{t_P} = 2^{-N}$ Unit

$5.39106 \times 10^{-44}(s) = 2^{-N} \times 18.8 \times 10^{9} \times 3.1536 \times 10^{6}$ (s)

$2^{-N} = 9.093 \times 10^{-61}$

**N** = 200

It is possible that the physical universe is one of many creation events within Unity Screen; thus **N** could be significantly larger.

Practically, we can assume the clock speed of the MS that creates the physical universe is the maximum detectable frequency of electromagnetic waves in the physical universe.

According to this model, all phenomena, including electromagnetic waves, are a processing output of the MS. Therefore, the frequency of the processing output cannot exceed the clock speed of the MS.

In our physical universe, the highest measurable frequency of an electromagnetic wave is Gamma ray radiation that is at least $10^{19}$ Hz.

Thus,

$2^{(N-1)}$ cycles/Unit = $10^{19}$ cycle/Sec

$2^{(N-1)} / 18.8 \times 10^{9} \times 3.1536 \times 10^{6}$ (s) = $10^{19}$ /s

$2^{(N-1)} = 5.929 \times 10^{35}$

**N** = 119

Thus it can be concluded that the MS that constructed the physical universe operates with an **N** value of at least 119.

### *MIND*

Mind is a partition of Source Mind. The partitioning is a processing output of MS achieved by running program {Individuality} or {I} or {Self}. This program produces a sense of "I" or "self," and identifies itself with an individual EI.

Mind is the subjective aspect of entity.

As a partition of Source Mind, mind shares the same qualities and traits as Source Mind. Metaphorically, it can be likened to the fact that every droplet of water in the ocean has the same wetness as the ocean.

Therefore, mind has the power and capability of conception, perception, and self-awareness. Mind also has access to the three faculties of MS: data, program, and processor.

As each individual EI is normally localized at a specific level of creation and specific space and time, mind has limited access to data, program, and computing capability.

As one aspect of entity, each mind is further partitioned into many lower minds at the subsequent level of creation. Mind, and its subsequent lower minds, computes using different MSs operating at different clock speeds. Each mind is also a partition of its higher mind.

A human mind operating at the physical level conceives the virtual entities by programming a physical computer. The virtual entities, however, cannot perceive the processing output displayed on the computer screen.

Likewise, the higher mind conceives the physical entities by programing a MS at a higher level creation. The human mind is, however, unlike the virtual reality game entities, able to perceive the physical world displayed in 3D space as objective existence and thus able to experience an individual, localized personal life.

Therefore, higher mind conceives the data and programs in the MS at a higher level creation; lower mind perceives and experiences the processing output of the MS at a lower level creation.

### *HUMAN MIND*

The human mind shares the same qualities and attributes of its higher mind and, ultimately, that of Source Mind. It has the power and capability of conception, perception, and self-awareness.

A human mind is associated with a human body, including the brain. Our physical body is localized at the physical level and in specific physical space and time. This imposes limitations on our access to data and programs.

Each individual human mind perceives an individual world that is a processing output determined by its access to data and programs. On our planet there are approximately seven billion worlds perceived by seven billion human minds. Two individual worlds can only be identical if the two individual human minds process the same data with the same programs.

The content of a human mind is the processing output of the MS displayed in space and in the body.

Space is used as a display onto which the EI's visual output is projected.

The brain is used as a display onto which thoughts, feelings, and emotions are projected.

The physical body is used as a display onto which bodily sensations and actions are projected.

The development of the human body, including the brain, is a process of upgrading the display so that it can display the output of MS from accessing increasing amounts of data and running an increasing number of programs with increasing complexity. This allows for the expansion of life experiences of the human mind.

At a particular moment during the early stage of our lives, each human mind starts to access and run program {Time}. The moment this happens is the personalized PM for that human being.

### *RELATIVITY OF REALITY*

Reality is what is perceived by the mind as objective existence independent of processing.

A human mind operating at the physical level creation can conceive a physical computation system. A human mind can also conceive a virtual world by programming a physical computer and perceives the processing output displayed on the screen.

Likewise, higher mind can conceive space and the physical world by programing a MS at a higher level creation.

From the perspective of the higher mind, the physical level existence is the processing output of the MS and therefore is a processed existence.

Physical object is projected into space as an output of the MS in the form of 3D poxel barcode arranged in specific configurations and patterns. It can be said that poxels are the building blocks of matter in the physical universe.

From the perspective of the human mind, however, the perceived physical world is an objective existence.

The fact that the physical world is perceived by the human mind as physical reality is due to the availability of the abundant resources in the MS, including the following:

- Large memory and processing capability.
- Display being a 3D space with high resolution.
- Programs that give physical properties to objects such as {Transparency}, {Solidity}, {Rigidity}, {Mass}, {Color}, {Texture}, etc.
- Programs that govern the behaviors of physical objects and their interactions, such as {Laws of Nature}, {Gravity}, {Field}, {Force}, {Electromagnetism}, {Mechanics}, {Energy}, etc.
- Complexity of the human brain that is capable of displaying a wide range of physical properties and concepts as complex electrical and chemical signal patterns.

When a human mind processes {Space}, a 3D grid with regularly arranged alternating poxels and voids are projected. Poxels are programed to be transparent so space appears to be empty.

When a human perceives an object in space, for example, an apple, the 3D poxel barcode dataset is scanned by the eyes to trigger the execution of program {Apple}. This produces a templet "Apple-ness" followed by adding more details and properties such as color and texture in the brain. The 3D image of an apple is then projected into space by the human eyes. An apple EI in a specific location in space defined by the dataset is thus perceived by the human mind, as illustrated in Figure 13.

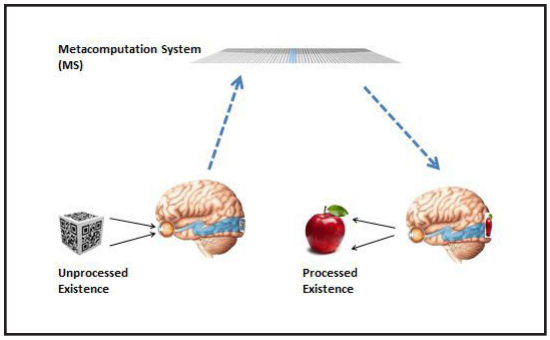


**Figure 13.** Perception of an apple in space. Data needs to be processed before a meaningful object can be perceived.

Programs such as {Mass} and {Gravity} ensure that the apple EI falls to the ground when it is detached from the tree branch. Programs such as {Solidity} and {Rigidity} ensure that the apple EI stays on top of the surface of the ground and doesn't go through the earth EI.

Our higher minds program the physical world. Some of these programs give processing outputs expressed as mathematical laws, scientific theories, laws of nature, arts, technologies, and industrial processes such as energy generation, product design, development, manufacturing, and application. Programs that are robust, reliable, and repeatable are accepted as mainstream programs at certain periods of time in human history.

In theory, mainstream programs can be interrupted or altered by the higher mind to cause phenomena that appear to violate and disrupt the physical laws of nature. Nevertheless, at our physical level of existence, miracles and paranormal phenomena are rare, generally nonrepeatable and uncontrollable. They only occur in some special circumstances.

### *FURTHER RESEARCH*

Further research is needed to discover programs that compute not only EI's geometric properties but also physical properties such as {Transparency}, {Solidity}, {Rigidity}, {Color}, etc.

{Laws of nature} governing the behaviors of physical objects and their interactions, involving {Mass}, {Energy}, {Force}, {Gravity}, {Field}, {Electromagnetism}, {Mechanics}, {Heat}, etc, should be determined.

Other challenging tasks include the discovery of programs that can compute the full range of human experiences including thoughts, feelings, emotions, sensations, and actions.

Ultimately, we will be able to write every word and sentence in human languages with codes.

Metacomputics is the systematic study of the origin, fundamental structure, composition, nature, properties, dynamics, and applications of the MS that constructs and operates the universes as its processing output.

## SUMMARY

The Metacomputics model is proposed to support the hypothesis that the physical universe is the processing output of computation.

Proposed Metacomputics model assumes the existence of an operating computer in Platonic realm.

Platonic computer is derived from a three-tier hierarchy construct of Platonic objects and it consists of three faculties: data, program, and processor.

The Metacomputation system (MS) is made by, of, with, from Consciousness.

The MS is the unprocessed existence of creation. The processing output of the MS is the processed existence of creation.

The model is developed from the convergence of metaphysics and computational theories. It offers a new perspective and clarity on many important concepts and phenomena that have perplexed humans for millennia, including consciousness, existence, creation, reality, time, space, multiverse, laws of nature, language, entity, mind, experience, thought, feeling, emotion, sensation, and action.

According to this model, the following can be deduced:

- Time is one-directional perpetual progression of a pixel square wave in the MS Grid that completes one switching cycle in Present Moment.
- Present Moment is the temporal moment when switching, and, therefore, computation takes place.
- Poxels are the 3D expression of the power of Source Mind in space.
- Poxels are the fundamental building blocks of the physical universe.
- Space is constructed with alternating regularly patterned poxels and voids in a 3D grid.
- Space is a 3D display onto which processing output of the MS is projected.
- Many levels of creation are in existence. Each level of creation is constructed from different MSs operating at different clock speeds.
- The physical universe is one of many parallel universes.
- Time dilates when ascending from lower to higher levels of creation.
- The MS that constructs the physical universe has at least $4^{119}$ bits memory.

The following can be implied:

- Words are created to name programs. The true meaning of a word is the perception experienced by the mind from executing the program.
- An entity is a being with both subjective and objective aspects. The objective aspect of an entity is the processing output of MS displayed in space as a 3D image. The subjective aspect of an entity is its mind.
- A physical entity exists as different entity images at different levels of creation.
- All entity images run program {life cycle} that progresses them through the stages of inception, expansion, deterioration, and termination in time.
- A mind is a partition of its higher mind and ultimately a partition of Source Mind.
- A mind and its subsequent lower minds compute using different MSs operating at different clock speeds.
- Entity images are generated in the MS and projected into space by the sense organs. Physical eyes are projectors as well as receptors.
- The brain is a display onto which thoughts, feelings, and emotions are projected as complex electrical and chemical signal patterns that can be experienced by the mind.
- Higher mind conceives the data and programs in the MS at a higher level creation; lower mind perceives and experiences the processing output of the MS at a lower level creation.

### ACKNOWLEDGEMENT

The author would like to thank all those who have contributed to the development of computation theories and technologies that have provided conceptual tools for this work.

Many great minds and their thoughts also provided a rich source of inspiration for this work. These include the following:

- Laozi's "Dao gives birth to One, One gives birth to Two, Two give birth to Three, Three give birth to everything";
- Parmenides's "The Unchanging One";
- Heraclitus's "The succession of opposites as a base for change" and "Permanent flux";
- Hegel's "three-valued logical model";
- Plato's "allegory of the cave" and "Realm of Forms";
- Pythagoras's "number as essence of Universe";
- Kant's "un-removable time-tinted and causation-tinted sunglasses";
- Locke's "blank canvas mind";
- Berkeley's "to be is to be perceived."

### REFERENCES

Bolognesi, T. "Algorithmic Causal Sets for a Computational Spacetime." In *A Computable Universe: Understanding and Exploring Nature as Computation*, edited by H. Zenil, 451–78. World Scientific Publishing, 2012.

Chaitin, G. "Life as Evolving Software." In *A Computable Universe: Understanding and Exploring Nature as Computation*, edited by H. Zenil, 277–302. World Scientific Publishing, 2012.

Deutsch, D. *The Fabric of Reality*. Penguin Press: Allen Lane, 1997.

Fredkin, E. "Finite Nature." *Proceedings* of the XXVIIth Rencotre de Moriond, 1992.

Fredkin, E. "A Computing Architecture for Physics." In Computing Frontiers, 273–79. Ischia: ACM, 2005.

Hooft, G. 't. "Quantum Gravity as a Dissipative Deterministic System." *Class. Quant. Grav.* 16 (1999): 3263–79. http://arxiv.org/abs/gr-qc/9903084.

Hutter, M. "The Subjective Computable Universe." In *A Computable Universe: Understanding and Exploring Nature as Computation*, edited by H. Zenil, 399–416. World Scientific Publishing, 2012.

Lloyd, S. "The Computational Universe: Quantum Gravity from Quantum Computation." *Quantum Physics* (2005). http://arxiv.org/abs/quant-ph/0501135.

Schmidhuber, J. "A Computer Scientist's View of Life, the Universe, and Everything." In *Foundations of Computer Science: Potential – Theory – Cognition, Lecture Notes in Computer Science*, edited by C. Freksa, 201–08. Springer, 1997.

Tegmark, M. "The Mathematical Universe." In *Visions of Discovery: Shedding New Light on Physics and Cosmology*, edited by R. Chiao. Cambridge: Cambridge University Press, 2007.

Weizsäcker, ^ von, Friedrich, Carl. *The Unity of Nature*. New York: Farrar, Straus, and Giroux, 1980.

Wheeler, John A. "Information, Physics, Quantum: The Search for Links. In *Complexity, Entropy, and the Physics of Information*, edited by W. Zurek (Redwood City, California: Addison-Wesley, 1990).

Whitworth, B. "Simulating Space and Time." *Prespacetime Journal* 1, no. 2 (March 2010).

Wolfram, S. "A New Kind of Science." Wolfram Media, 2002.

Zizzi, P. "Spacetime at the Planck Scale: The Quantum Computer View." 2005. http://arxiv.org/abs/gr-qc/0304032.

Zenil, H. "Introducing the Computable Universe." In *A Computable Universe: Understanding and Exploring Nature as Computation*, edited by H. Zenil. World Scientific Publishing, 2012

Zuse, K. *Calculating Space*. Cambridge, MA: MIT, 1969.

# *Toward a Philosophy of the Internet*

László Ropolyi
**EÖTVÖS UNIVERSITY, BUDAPEST, HUNGARY**

The appearance and the extended use of the internet can probably be considered as the most significant development of the twentieth century. However, this becomes evident if and only if the internet is not simply conceived as a network of interconnected computers or a new communication tool, but as a new, highly complex artificial being with a mostly unknown nature. An unavoidable task of our age is to use, shape, and, in general, discover it—and to interpret our praxis, to study and understand the internet, including all the things, relations, and processes contributing to its nature and use.

Studying the question what the internet is and its history—apparently—provides a praxis-oriented answer.[1] Based on the social and cultural demands of the 1960s, networks of interconnected computers were built up, and in the 1980s a worldwide network of computers, the net, emerged and became widely used. From the 1990s the network of web pages, the world wide web, has been built on the net. Using the possibilities provided by the coexisting net and web, social networks (such as Facebook) have been created since the 2000s. Nowadays, networking of connected physical vehicles, the emergence of the internet of things, the *IoT*, seems to be an essential new development. Besides these networks there is a regularly renewed activity to form sharing networks to share "contents" (files, material and intellectual property, products, knowledge, services, events, human abilities, etc.) using, e.g., streaming or peer-to-peer technologies. In this way, currently, from a practical point of view, the internet can essentially be identified as a complex being formed from five kinds of intertwined coexisting networks: the net, the web, the social networks, the IoT, and the sharing networks.

Furthermore, as it is easy to see, especially in the case of social and sharing networks, the internet cannot be identified and its development cannot be understood independently from the historical-societal and cultural environment in which it is launched and used. Identifying shaping influences of certain social and cultural relationships on the formation of the internet makes it easier for us to consider and identify the opposite relationships—i.e., to study the social and cultural impacts of internet use. In other words, accepting the idea of the social construction of the internet as a technology can help us understand the social and cultural consequences of its use.[2] Thus, it seems to be useful to employ a social and cultural context in the examination of the nature of the internet.

Taking into consideration the praxis of internet use, its two important characteristics come into sight. First, it is obvious enough that the mode of internet use changes very quickly and in an almost unpredictable way. The reasons for this course of events can be associated with the second characteristic of internet use: internet users are typically not just passive acceptors of the rules of use prescribed by the constructors of a given internet praxis, but they are active agents.[3] In fact, in the case of the internet, the constructor and user roles typically interlock with each other.

In this way, in order to identify the very nature of the internet and its characteristics, we have to understand the emergence and formation of a complex of several intertwined coexisting and interacting networks shaped by experts and active users in the changing social and cultural environments of the late Modern Age. Over and above, we have to disclose and consider the social and cultural impacts of this complex being, and to study the meaning of the construction of the internet and that of the ubiquity of its human use.

## METHODOLOGICAL CONSIDERATIONS—TRENDS IN INTERNET RESEARCH

Confronting these intellectual challenges, research on the internet had already been initiated practically at the time of the emergence of the internet. In the beginning, most research was performed in the context of informatics, computer sciences, (social) cybernetics, information sciences, and information society, but from the 1990s a more specific research field, "internet research," started to form, incorporating additional ideas and methodologies from communication-, media-, social-, and human sciences. From the 2000s, internet research can be considered as an almost established new (trans-, inter-, or multidisciplinary) research field.[4]

It is not surprising at all that the new discipline faced serious methodological difficulties. Besides its trans-, inter-, or multidisciplinary ambitions, internet research is also shaped by the following additional circumstances:

i) The *historical*, *social*, and *cultural* context of the emergence and deployment of the internet. Elaboration of the basic principles of internet construction and the realization of these plans fundamentally take place in the late modern or postmodern age, in the second half of the twentieth century, in a parallel trajectory with becoming widespread and achieving a cultural dominancy of the postmodern values and ideology.[5] Postmodern ideology is not shaped by (modern) sciences; it has a rather technological, more precisely, *techno-scientific*, background and preference. This way it is easier to understand postmodern constructions in a technological or a techno-scientific context.

ii) The "omnipresence" or *ubiquity* of the internet. Our experiences in connection with the internet are extremely diverse in quality and infinitely extended in quantity. The fact that the internet can be found in and has an impact on the whole human practice is a source of many methodological difficulties: findings of any meaningful abstractions about the internet, identification of real causal relationships, recognition of the borders of beings in an extended continuum, interpretation of the social and cultural effects of the internet, etc., are extremely difficult. The internet as a research object is a highly complex organization of numerous problematically identifiable complex entities.[6]

iii) A further difficulty is the essential *simultaneity* of the processes and their analyses, which means that the hard problems of participant observation will necessarily be present in the research procedure.

In response to these ambitions and difficulties, four different approaches to internet research have emerged in the last two decades:

a) *Modern scientific approach*. In this kind of research, the main deal is accepting the validity of an established (modern) scientific *discipline* to apply its methodology on the internet and internet use. An aspect of the internet or internet use is considered as a subject matter of the given science.[7] In this way the internet or internet use can—at best—be described from computational, information technological, sociological, psychological, historical, anthropological, cognitive, etc., points of view. This is a very popular praxis; however, such research is necessarily insensitive to the characteristics of the subject matter outside of their disciplinary fields due to the conceptual apparatus and the methodology of the selected scientific discipline, in this case to the specificity of the internet and internet use. Outcomes of these studies can be considered as specific (internet-related) disciplinary statements of which the significance on the specificity of the internet is not obvious at all.

When researchers in these disciplines consider one or another thing as an interesting aspect of the internet, their choice is more or less "evident"—i.e., it is a pragmatic presupposition on the internet. In this way it is almost impossible to see the significance of the given aspect of the internet (and the given disciplinary approach) in the understanding of the internet. Without careful philosophical analysis on the nature of the internet, it is not trivial at all how relevant sociology, psychology, informatics, anthropology, or any other classical scientific discipline relates to its description.

Additionally, in this methodology the inter-, trans-, or multidisciplinarity aspect of internet research is fulfilled in an indirect way: the big set of traditional scientific descriptions of the internet includes items from many different, but usually unrelated, disciplines. Taking into account some considerations of the philosophy of science, coexisting disciplines and their joint application to the fundamental conditions of the internet can perhaps produce much more coherent outcomes.

b) *Postmodern studies approach*: elaborating and applying a pluralist postmodern methodology of the so-called *studies*. Studies include concrete, but case by case potentially different mixtures of disciplinary concepts and methodologies that are being applied to describe the selected topic. Application of studies (e.g., internet studies, cultural studies, social studies, etc.) methodology results in the creation of a huge number of relevant but separated and necessarily unrelated facts. Most research published in studies are well informed on the specificities of the internet, so the selected methodological versions in the different studies can fit well to a specific characteristic of the internet or internet use, but the methodological plurality of the different studies prevents reaching any generalized, universally valid knowledge of the internet. Nowadays most internet research is performed in this style. Collections of studies[8] and articles in online and offline journals devoted to internet research (*First Monday*, *Journal of Computer-Mediated Communication*, *Internet Research*, *Information Communication and Society*, *New Media & Society*, etc.) can be considered as illustrative examples.

c) Internet *science* approach to the internet and/or internet use. Among researchers of the internet, there is a lack of consensus regarding how to best describe the internet theoretically, i.e., whether it is a (scientific) *theory* or rather a *philosophy* of the internet that is needed. Scientific theories on the internet presuppose that the internet is an independent entity of our world and seek for its specific theoretical understanding and description. Because of the complexity of the internet, it is not surprising that comparing these theories to the classical scientific theories have a definite trans-, inter-, or multidisciplinary character. They usually combine the methodological and conceptual apparatus of social-scientific (sociology, psychology, political theory, law, political economy, anthropology, etc.), scientific, mathematical, and engineering (theory of networks, theory of information, computing, etc.) disciplines to create a proper "internet scientific" conceptual framework and methodology. Some of these theories really fit into a recent scientific standard providing universally valid knowledge in the form of justifiable or refutable statements, with empirical background and philosophical foundations. Their empirical background frequently includes the above mentioned disciplinary or

studies-origin facts, and their philosophical foundations vary case by case.

Although attempts to craft an internet theory has been observable from a relatively early phase of the formation of the internet[9] the whole history of theorizing the internet is very short, so it is not surprising that there is no universally accepted theory. Based on their different theoretical/philosophical presuppositions on the fundamental specificity of the internet, recently Tsatsou identified three characteristic groups of theories.[10] In these groups of theories, the specificities of the internet are determined by (i) its technologically constructed social embeddedness, or (ii) the specific political economy of its functioning, or (iii) the formation of specific networks. In this way the internet is (i) a social entity, which is fundamentally technologically constructed, or (ii) a social entity which necessarily participates in the reproduction of social being, or (iii) a particularly organized mode of social being.[11]

The diversity of these typical theoretical approaches casts light on the shortage of internet science: there is no consensus about the fundamental specificities of the internet. In other words, the philosophical foundations of internet science, the foundational principles on the nature of the internet, are essentially diverse ones—and in many cases they are naïve, unconsciously accepted, non-reflective, uncertain, or vague presuppositions. Philosophical considerations on the nature of the internet and on the effective principles of internet science can usefully contribute to overcoming these difficulties.

This situation is practically the same as we have (or had) in cases of any kind of sciences: the subject matter and the foundational principles of a scientific discipline are coming from philosophical considerations. As an illustration we can recall the determining role of natural philosophy in the formation of natural sciences, or the role of philosophy of science in the self-consciousness functioning of any developed scientific disciplines.

However, scientific theories of the internet face additional difficulties if they want to reflect on the (pluralistic) postmodern characteristics of the internet, on the quick and radical changes in internet use, on the extreme complexity of this being, and on the necessary presence of participant observation. Recently, there is a better chance of producing acceptable treatments of these difficulties in philosophies than in sciences.

d) *Philosophy of the Internet* approach. Like the internet science, philosophy of the internet also provides a theoretical description of the internet, but it is a completely different theoretical construction—at least if we do not identify philosophy with a kind of linguistic-logic attraction, but we see it traditionally as the conceptual reconstruction of our whole world set up by critical thinking.

As Aristotle declared in his *Metaphysics*, there are two kinds of theoretical methodologies: the scientific disciplines describe beings from a selected aspect of them, but philosophy describes "beings as beings," as a whole, considering them from all of their existing aspects. In this tradition, focusing on a given being, discovering and disclosing all of its interrelations of everything else, and in this way, characterizing the being from all of its aspects, the philosopher builds up a complete world in which the given being exists. Philosophical understanding is proceeding on the parallel "constructions" of the "being as being" and the "whole" world.[12] An ontology created in this way is essentially different from the ontologies constructed in computer sciences. Currently, this Aristotelian style of making philosophy is not really fashionable, and, in fact, not so easy to perform, but it seems to be not impossible and perhaps even necessary if one wants to understand a new kind of being of our recent word, as the internet is.

So the crucial distinction between sciences and philosophy makes clear the different possibilities of science and philosophy in the theoretical description of the internet.[13] Considering further the science-philosophy relationships, it becomes obvious that there is no science without philosophy. Historically, (European) philosophy emerged several hundred years before science did; science does not exist without (or prior to) philosophy. Of course, this is absolutely true in case of any concrete disciplines: emerging scientific disciplines are based on and spring out from philosophical (e.g., natural-philosophical) considerations and they include, incorporate, and develop these contents further. What is a natural object? What is a living organism? What is a constitution? And how can we identify and describe their nature and characteristics? Any scientific understanding presupposes such conceptual constructions. However, these procedures sometimes remain hidden, and the given scientific activity runs in an unconscious manner. These situations provide possibilities for the philosophy of science to clarify the real cognitive structures.

Following these intellectual traditions, if we want to construct an internet science, we need some kind of philosophical understanding of the internet prior to the scientific one. What is the internet? What are its most fundamental specificities and characteristics? What are the interrelationships between the internet and all the other beings of our world? Only the philosophical analyses can provide an understanding of *the* internet *as the* internet, a theoretical description of its very nature, as a totality of its all aspects, as a whole entity.

These are the reasons that I have proposed for building a philosophy of the internet prior to the scientific theory of it.[14] First of all, taking into account the huge amount of its aspects, appearances, modes of use, etc., we should have to understand the nature of the internet and to suggest useful concepts, valid principles, and operable practices for its description. I have proposed to construct a philosophy of the internet in an analog manner as the *philosophy of nature* (or natural philosophy) was created before (natural) sciences.

However, besides this possibility, there are additional possibilities to contribute to the philosophy of the internet. Realizing the crucial social and cultural impacts of internet use, philosophers have started to consider the influence of internet use on philosophy.[15] Typically, they focus on

a particular aspect or side of the internet or internet use and put it into a philosophical context. In this way—doing research on the "philosophical problems of the internet"—one can identify the philosophical consequences of some kind of specificity of the internet or can disclose something on the nature of the specificity of the internet. This is the philosophy of the internet making in an analog manner as we used to make research in the *philosophy of science* or philosophy of language, or philosophy of technology, etc.

In the case of the natural philosophical type of the philosophy of the internet, we should have to create a complete philosophy in order to propose an understanding of the internet in our world, and an understanding of our world which includes the internet. In case of the philosophy of science type of the philosophy of the internet, we should have to apply, improve, or modify an existing philosophy in a sense in order to propose an understanding of a philosophical problem of the internet, and an understanding of a philosophical problem created by the existence and use of the internet. The latter type of philosophy is closer to internet science, while the former approach is closer to a real philosophy of the internet.

As I see it, the so-called philosophy of the Web (Philoweb) initiative is a representative of the "philosophical problems of the internet" type of research.[16] The typical analyses in their papers focus on a particular aspect of the internet (or the web) or focus on particular philosophical approaches (e.g., semantics, ontology) and try to conclude several consequences in these contexts.

Another important work in a similar philosophical methodology is provided by Floridi.[17] Floridi's philosophical works, for example, describe the changing meanings of several classical philosophical concepts (like reality) because of the extended internet use and vice versa: internet use is taking place in a non-traditional reality.

Some additional philosophical approaches focus on more specific disciplines (e.g., computer-mediated communication,[18] ethics[19]) or problems (e.g., embodiment,[20] critical theory of technology[21]).

Summing up, the philosophy of the internet can be considered as a new field of culture, a recent version of philosophizing with the ambitions to build philosophies in the era of the emergence and deployment of the internet and internet use, and taking these new circumstances seriously. It necessarily has different realizations, with different ideologies, values, emphases, cognitive structures, languages, accepted traditions, etc. There are at least two metaphilosophical attitudes toward this new cultural entity: a) creating an original version of philosophy, taking into consideration all of the experiences in the era, b) modifying existing philosophical concepts, systems, approaches, and meanings in order to understand the emerging problems of the internet era.

## SPECIFICITIES OF AN "ARISTOTELIAN" PHILOSOPHY OF THE INTERNET

In the last ten to fifteen years, I have developed a natural philosophical type of the philosophy of the Internet which I call "*Aristotelian" philosophy of the Internet*. As an illustration of the above mentioned ambitions, now I will try to sum up its main ideas.

This philosophy of the internet has Aristotelian characteristics in the following sense:

a) It is clear from the history of (natural) sciences that natural philosophy has a priority to any kind of natural sciences. The most successful natural philosophy (or philosophy of nature) was created by Aristotle. In his thinking, a "division of labor" between philosophy and sciences was clearly declared: understanding the being as being, or understanding an aspect of a being. Historically and logically, in the first step we can "philosophically" understand a given being and its most essential characteristics, and in a second step, based on this knowledge, we can create a science for their further understanding. In the case of the internet, first we try to understand its nature and its most fundamental characteristics "philosophically," and in the second step, an internet science can be created based on this knowledge.

b) In the Aristotelian view, beings (and the world as well) have a complex nature, and for their understanding we have to find a complex methodology. His crucial tool for this purpose was his causal "theory": everything has four interrelated, but clearly separated, causes—the material, the formal, the efficient, and the final cause. Applying this version of causality, the complex nature of any beings (and the world) can be disclosed. In the case of the internet (as a highly complex network of complex networks) this is a very important possibility for a deeper understanding. Of course, the concrete causal contexts will be different related to the original Aristotelian ones, so we will use the technological, the communication, the cultural, and the organization contexts to describe the highly complex nature of the internet.

c) There are several additional, but perhaps less crucial, Aristotelian components in my philosophy of the internet. Aristotle made a sharp distinction between natural and artificial beings (especially in his *Physics*). Based on this distinction, the fundamental role of technologies—as creators of the artificial spheres of beings—in the human world is really crucial, so I tried to find a technological (or techno-scientific) implementation for all of the aspects of the internet. Moreover, in the "solution" of several classical philosophical problems, I followed the Aristotelian traditions—e.g., my interpretation of virtuality (which is an important task in this philosophy of the internet) is based on the Aristotelian ontology.[22]

It is clear at first glance that the internet is an artificial being created mainly from other artificial beings. This means that its philosophical understanding is necessarily based on the philosophical understanding of other beings, so it has necessarily a kind of "metaphilosophical" characteristic.[23] The general view of the Aristotelian causality (in

the above mentioned way) can be considered as a metaphilosophical tool, which presupposes to understand and use philosophies of technology, philosophies of communication, philosophies of culture, and philosophies of organization for producing a complex philosophy of the internet. Additionally, it is useful to study and use the philosophical views on information, reality and virtuality, community, system and network, modern and postmodern, knowledge, human nature, spheres of human being, etc., in the process of constructing the philosophy of the internet.

As is clear from the statements above, this philosophy of the internet is not just about an abstract description of the internet, since it is included in and coexists with natural, human, social, and cultural entities in a complex human world. According to our research strategy, first, we examine the complex *nature* of the internet, and then we analyze the social and cultural *impacts* of its use. The two topics are, of course, closely related. The interpretability of social and cultural effects, to be discussed in the second step, requires a kind of understanding of its nature in which social and cultural effects are conceivable at all. In certain cases, this involves trying to make use of connections which are uncommon in the task of interpreting the internet. Thus, for example, we engage in discussions of philosophy, philosophy of technology, communication theory, epistemology, cognitive science, and social and cultural history instead of directly discussing the internet in "itself."

Taking into consideration the social and cultural factors which *define* or *shape* the nature of the internet obviously helps identify those social and cultural *effects that occur* in the course of internet use.

## ON THE NATURE OF THE INTERNET

In the "natural philosophical type" or the Aristotelian philosophy of the internet, the main task is to understand the nature of the internet and some of its essential characteristics. Below, a short outline of the components of this philosophy is presented in the form of theses.[24]

In the Aristotelian philosophy of the internet, we conceive of the internet in *four*—easily distinguishable, but obviously connected—*contexts*: we regard it as a system of *technology*, as an element of *communication*, as a *cultural* medium, and as an independent *organism*.

1) *Technological context*. I propose that we conceive of technology as a specific form or aspect of human agency, the realization of human control over a technological situation. In consequence of the deployment of this human agency, the course and the outcome of the situation seem no longer governed by natural constraints but by specific human goals. Human control of technological situations yields artificial beings as outcomes. With the use of technology, man can create and maintain artificial entities and, as a matter of fact, an artificial world: its own "not naturally given" world and she/he shapes her/his own nature through her/his own activity. Every technology is value-laden—i.e., technologies are not neutral; they unavoidably express, realize, and distribute their built-in values during usage. The internet obviously is a technological product, and at the same time it is a consciously created technological system, so, like other technologies, the internet also serves human control over given situations.

However, the internet is a specific system of technology; it is an *information* technological system. It works with information rather than with macroscopic physical entities. As I see it, information is created through interpretation, so a certain kind of hermeneutical practice is a decisive component of information technologies. In consequence, information—and all kinds of information "products"—is virtual by nature. Though it seems *as if* it was real, its reality has a certain limited, finite degree.[25]

The information technological system of the internet—in fact, we can talk about a particular type of system, that is, network—consists of computers which are interconnected and operated in a way which secures the freedom of information of the individuals connected to the network: the control over information about themselves and their own world in space, time, and context.

Thus, from a technological point of view, the internet is an *artificially created* and maintained *virtual sphere*, for the operation of which the functioning of the computers connected into the network and the concrete practices of people's interpretations are equally indispensable.

2) *Communication context*. For the characterization of the internet as an element of communication, we can understand communication as a certain type of technology, the goal of which is to create and maintain communities. Consequently, the technologies of communication used on the internet are those technologies with the help of which particular—virtual, open, extended, online, etc.—communities can be built. The individual relationships to the communities that can be built and the nature of the communities can be completely controlled through technologies of the internet (e-mail, chat, lists, blogs, podcast, social networks, etc.). Communication through the internet has a network nature (it is realized in a distributive system); it uses different types of media, but it is a technology which follows a basically visual logic.

Thus, as regards communication, the internet is the *network of* consciously created and maintained extended plural *communities*, for the functioning of which the harmonized functioning of computers connected to the network as well as the individual's control over his own communicative situations are needed.

3) *Cultural context*. From a cultural point of view, the internet is a medium which can accommodate, present, and preserve the wholeness of human culture—both as regards quality and quantity. It can both represent a whole cultural universe and different, infinitely varied cultural universes (worlds).

Culture is the system of values present in coexisting communities; it is *"the world of" communities*. Culture is the technology of world creation. Culture shapes and also expresses the characteristic contents of a given social system. Each social system can be described as the

coexistence of human communities and the cultures they develop and follow. Schematically,

**society = communities + cultures**

The individual is determined by her participation in communities and cultures, as well as his contribution to them.

The internet accommodates the values of the late modern age, or the "end" of modernity. That is, it houses late modern worlds. Late modern culture contains modern values as well, but it refuses their exclusivity and it favors a plural, postmodern system of values. The way of producing culture is essentially transformed: the dichotomy of experts creating traditional culture and the laymen consuming it are replaced by the "democratic nature" of cyber culture: each individual produces and consumes at the same time.

Thus, from a cultural point of view, the internet is a network of virtual human communities, artificially created by man unsatisfied by the world of modernity; it is a network in which a postmodern system of values based on the individual freedom and independence of cyberculture prevails.

4) *Organism context*. From an organizational point of view, the internet is a relatively independent organism, which develops according to the conditions of its existence and the requirements of the age. It is a (super)organism created by the continuous activity of people, the existence, identity, and integrity of which is unquestionable; systems, networks, and worlds penetrating each other are interwoven in it. It has its own, unpredictable evolution: it develops according to the evolutionary logic of creation and human being, wishing to control its functioning, is both a part and a creator of the organism.

The indispensable vehicles are *the net*, built of physically connected computers, *the web*, stretching upon the links which connect the content of the websites into a virtual network, *the human communities* virtually present on the websites organized into social networks, the interlinked human things as well as the infinite variations of individual and social *cultural entities and cultural universes* penetrating each other.

The worldwide organism of the internet is imbued with values: its existence and functioning constantly creates and sustains a particular system of values: the network of postmodern values. The non-hierarchically organized value sphere of *virtuality*, *plurality*, *fragmentation*, included *modernity*, *individuality*, and *opposition to power*, interconnected through weak bonds, it penetrates all activity on the internet—moreover, it does so independently of our intentions, through mechanisms built into the functioning of the organism.

Thus, from the organizational point of view, the internet is a superorganism made of systems, networks, and cultural universes. Its development is shaped by the desire of late modern man to "create a home," entering into the network of virtual connections impregnated with the postmodern values of cyberculture. For human beings, the internet is a new—more homely—sphere of existence; it is the exclusive vehicle of web-life. Web-life is created through the transformation of "traditional" communities of society and the cultures prevailing in the communities. Schematically, web-life = "online" communities + cybercultures.

To sum up, the internet is the medium of a new form of existence created by late modern man, a form that is built on earlier (i.e., natural and social) spheres of existence, and yet it is markedly different from them. We call this newly formed existence *web-life*, and our goal is to understand its characteristics.

## SOCIAL AND CULTURAL IMPACT OF INTERNET USE

Based on this understanding of the internet, the social and cultural consequences of the internet use can be disclosed and characterized as crucial characteristics of the web-life. The following two analog historic-cultural situations (analogies can provide a useful orientation within a highly complex and fundamentally unknown situation) can be tackled in the hope of obtaining a deeper understanding of the impact of the internet use on our age:

1) *The Reformation of Knowledge*. For the study of the mostly unknown relations of web-life, it seems to be useful to examine the nature of *knowledge*, which was transformed as a consequence of internet use, its social status, and some consequences of the changes.

Inhabitants of the fifteenth and sixteenth centuries and of our age have to face similar challenges: citizens of the Middle Ages and modern "web citizens" or "netizens" participate in analogous processes. The crisis of religious faith unfolded in the late Middle Ages and in our age, the crisis of rational knowledge can be observed. In those times, after the crisis—with the effective support of reformation movements—we could experience the rise of rational thinking and the new, scientific worldview; in our times, five hundred years later, this scientific worldview itself is eventually in a crisis.

The *reformation of religious faith* was a development which evolved from the crisis of religious faith. The *reformation of knowledge* is a series of changes originating from the crisis of rational knowledge.

The scenes of the *reformation of religious faith* were religious institutions (churches, monasteries, the Bible, etc.). Nowadays, the *reformation of knowledge* is being generated in the institutional system of science: research centers, universities, libraries, and publishers.

In both cases, the (religious and academic) institutional system and the expert bodies (the structure of the church and the schools and especially universities, research centers, libraries, and publishers, as well as priests and researchers, teachers, and editors) lose their decisive role in matters of faith as well as science. The reformation of faith, ignoring the influence of ecclesiastical institutions, aims for developing an immediate relationship between

the individual and God. The reformation of knowledge creates an immediate relationship between the individual and scientific knowledge.

It is well known that book printing played an important role in the reformation of faith. Books are "tools" which are in accordance with the system of values of the world undergoing modernization. They made it possible to experience and reform faith in a personal manner as a result of the fact that the modern book was capable of accommodating the system of values of the Middle Ages. (But the typical usage of the book as a modern "tool" is not this but rather the creation and study of modern narratives in a seemingly infinite number of variations.).

In a similar way internet use plays an important role in the reformation of knowledge. The internet developed and became widely prevalent simultaneously with the spreading of the postmodern point of view. It seems that the crisis of modernity created a "tool" that fits with its system of values. It grows strong partly because of this accordance; what is more, people develop it further. However, at the same time, this "tool," the internet, seems to be useful for pursuing forms of activities which are built on the postmodern world but transcend it and also for the search for the way out of the crisis. (Postmodern thinking was itself created and strengthened by the—more or less conscious—reflection about the circumstances of the crisis, as the eminent version of the philosophy of the crisis.)

On the internet, ideas can be presented and studied in a direct way, in essence, independently of the influence of the academic institutional system. There are no critics and referees on websites; everyone is responsible for his own ideas. The reformers diagnose the transformation of the whole human culture because of the internet use: the possibility of an immediate relationship between the individual and knowledge is gradually forcing back the power of the institutional system of abstract knowledge (universities, academies, research centers, hospitals, libraries, publishers) and its official experts (qualified scientists, teachers, doctors, editors). The following question emerges today: How can we get liberated from the power of the decontextualized, abstract rationality that rules life? In the emancipation process that leads out of the crisis of our days, the reformation of knowledge is happening, using the possibilities offered by the internet. We can observe the birth of the yet again liberated man on the internet, who, liberated from the medieval rule of abstract emotion, now also wants to rid himself of the yoke of modernist abstract reason. But his or her personality, system of values, and thinking are still unknown and essentially enigmatic for us.

The reformation of faith played a vital role in the development process of the modern individual: harmonizing divine predestination with free will secured the possibility of religious faith, making the development of masses of individuals in a religious framework possible and desirable.

However, the modern individual that developed this way, "losing his embeddedness" in a traditional, hierarchical world, finds herself in an environment which is alien, even hostile to him or her. As a consequence of such fear and desire for security, the pursuit of absolute power becomes his/her second nature; the modern individual is selfish.

Human being, participating in the reformation of knowledge (after the events that happened hundreds of years before) is forced again into yet another process of individuation. Operating his/her personal relationship to knowledge, a postmodern individual is in the process of becoming. The postmodern personality, liberated from the rule of the institutional system of modern knowledge, finds him/herself in an uncertain situation: she herself can decide in the question of scientific truth, but she cannot rely on anything for her decisions.

This leads to a very uncertain situation from an epistemological point of view. How can we tackle this problem? Back then, the modern individual eventually asked the help of reason and found solutions, e.g., the principle of rational egoism or the idea of the social contract. But what can the postmodern personality do? Should she follow perhaps some sort of post-selfish attitude? But what could be the content of this? Could it be perhaps some kind of plural or virtual egoism? The postmodern personality got rid of the rule of abstract reason, but it still seems that s/he has not yet found a more recent human capacity, the help of which s/he could use in order to resolve his/her epistemological uncertainty.

From a wider historical perspective, we can see that people in different ages tried to understand their environment and themselves and to continue living by relying on abstract human capacities that succeeded each other. People in primeval societies based their magical explanation of the world on the human will—and we managed to survive. After the will, the senses were in the mythical center of ancient culture—and the normal childhood of humankind passed, too. Medieval religious worldview was built by taking into consideration the dominance of emotions—and this ended, too, at some point. In the age of the glorious reason, it was the scientific worldview that served the reign of man (rarely woman)—until now.

Today, the trust in scientific worldview seems to be teetering; the age of the internet has come. However, the problem is that we cannot draw on yet another human capacity since we have already tried them all, at least once. But have we? Do we still have hidden resources? Or can we say goodbye, once and for all, to the usual abstractions, and a new phase of the evolution of humankind is waiting for us, which is happening in the realm of the concrete?

2) *Formation of Web-Life*. In order to study the mostly unknown context of web-life, it seems to be useful to examine the nature of *human existence*, transformed through internet use and the consequences of the changes. Social scientists like Castells (2000), Wellman and Haythornthweait (2002), or Fuchs (2008) often characterize the consequences of internet use as pure *social* changes, including all kinds of changes into social ones, and disregard the significance of more comprehensive changes. We would focus on the latter one.

While using the internet, all determining factors and identity-forming relations change, which had a role in the evolution of humankind from the animal kingdom and in the process of the development of society. We can identify tool use, language, consciousness, thought, as well as social relationships as the most decisive changes in the process of becoming human and in the formation of web-life that has developed as a result of internet use.

The simultaneous transformations of animal tool and language use, animal consciousness and thought, as well as social relationships and the series of interwoven changes led to the evolution of humans and to the development of culture and society. Nowadays, the robust changes in the same areas are also simultaneous. They point in one direction, intensifying each other, and induce an interconnected series of changes. The quantity of the changes affecting the circumstances of human existence results yet again in the qualitative transformation of the circumstances of existence: this is the process of the *development of web-life*.

The material circumstances of tool making and tool use lose their significance and the emphasis is now on the most essential part of the process: interpretation. A crucial part of tool making is the interpretation of an entity in a different context, as different from the given (such as natural entities), and in this "technological situation" its identification as a tool. During internet usage, individual interpretations play a central role in the process of creating and processing information on different levels and in the information technologies that are becoming dominant. At the same time, the material processes that provide the conditions of interpretation are, to a large extent, taken care of by machines. Hermeneutics takes the central role of energetics in the necessary human activity of reproducing human relations.

The human double- (and later multiple-) representation strategy developed from the simpler strategies of the representation characteristic of how wildlife led to language, consciousness, thought, and culture. Double representation (we can regard an entity both as "itself" and "something else" at the same time) is a basic procedure in all these processes—including tool making—and an indispensable condition of their occurrence. The use of the internet radically transforms the circumstances of interpretation. On the one hand, it creates a new medium of representation in which—as in some sort of global "mind"—the whole world of man is represented repeatedly. On the other hand, after the ages of orality and literacy, it makes possible basically for all people to produce and use in an intended way the visual representation of their own world as well. Virtuality and visuality are determining characteristics of representation. We are living in the process of the transformation of language, speech, reading and writing, memory and thought.

"Traditional" human culture is created through the reinterpretation of the relations "given by nature." It materializes through their perpetual transformation and it becomes a decisive factor in the prevailing social relations. The cybercultural practices of the citizens of the web are now directed at the reevaluation of social relations, and as a result of their activities a cyber-, web- or internet-cultural system of relations is formed, which is the decisive factor in the circumstances of web-life.

The basically naturally given communities of animal partnership were replaced by the human structure of communities, which was practically organized as a consequence of the tool-use-based indirect, and language-use-based direct communicative acts. However, the control over communicative situations can be monopolized by various agents; as a result, it is burdened with countless constraints. The nature of the communities that come into existence under these circumstances can become independent from the aspirations of the participants: various forms of alienation and inequality can be generated and reproduced in the communities. The citizen of the web who engages in communication reinterprets and transforms communicative situations; above all, he changes power relations in favor of the individual: the citizen of the web can have full powers over her/his own communicative situations.

## CONCLUSION

Philosophy of the internet discloses that human existence is being transformed. Its structure, many thousand years old, seems to be changing. Built on the natural and the social spheres of being, a third form of existence is emerging: web-life. Human being is now the citizen of three worlds, and his/her nature is being shaped by these three domains, i.e., by the relations of natural, social, and web-life. Our main concern is the study of web-life, which has developed as the result of internet use. From the position of the above proposed philosophy of the internet—besides illuminative cultural-historical analogies—the following cultural-philosophical topics seem to have fundamental significance in the understanding of the characteristics of web-life:

- The *knowledge* presented and conveyed through the internet valorizes the forms of knowledge which are characteristically situation-dependent, technological, and postmodern. The whole modern system of knowledge becomes reevaluated and, to a large extent, virtualized; the relationship to knowledge, reality, and truth takes a personal, concrete, open, and plural shape. The significance of the institutional system of science is diminished. Instead of scientific knowledge, technological or technoscientific knowledge and the technologies of interpreting knowledge are in the forefront.

- Besides *culture* that is created by the communities of society, individual cyberculture plays a more and more important role. The traditional separation of the producers and consumers of culture becomes more and more limited in this process. Supported effectively by information technologies, billions of the worlds of the citizens of web-life join the products of the professional creators of culture. Cyberspace is populated by the infinite number of simultaneous variations of our individual virtual worlds. Aesthetic culture gains ground at the expense of scientific

culture, and imagination becomes the human capacity that determines cultural activities.

- *Personality* becomes postmodern, that is, it becomes fully realized as an individual, virtually extremely extended, and acquires a playful character with ethereal features. A more vulnerable post-selfish web citizen is developed, compelled by a chaotic dynamics. Web citizens are mostly engaged in network tasks, that is, in building and maintaining their personalities and communities.

- Besides the natural and the social spheres, a sphere of web-life is built up. Now humans become the citizen of *three worlds*. The human essence moves towards web-life. The freedom of access to the separate spheres and the relationship of the spheres of existence are gradually transformed in a yet unforeseeable manner. Characteristics of web-life are shaped by continuous and necessarily hard ideological, cultural, political, legal, ethical, and economical conflicts with those of the traditional social sphere.

- Web-life as a form of existence is the realm of concrete existence. Stepping into web-life, the *"real history" of mankind* begins yet again; the transition from social existence to web-life existence leads from a realm of life based on abstract human capacities to a realm of life built on concrete capacities.

**NOTES**

1. See, e.g., Hobbes's Internet Timeline, 2018, https://www.zakon.org/robert/internet/timeline/; Living Internet, 2017, https://www.livinginternet.com/; History of the Internet, 2018, https://www.internetsociety.org/internet/history-internet/; etc.
2. The social construction of technology (SCOT) proposed by Bijker and Pinch ("The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other"; Bijker, Hughes, and Pinch, *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*) is a widely accepted view in the philosophy and sociology of technology and in the science and technology studies (STS).
3. Some relevant views can be found, e.g., in the literature of the so-called "user research." See, for example, Oudshoorn and Pinch, *How Users Matter. The Co-Construction of Users and Technologies*; or Lamb and Kling, "Reconceptualizing Users as Social Actors in Information Systems Research"; or in a more concrete, internet-related context see Feenberg and Friesen, *(Re)Inventing the Internet: Critical Case Studies*.
4. As an illustration: during the last fifteen to twenty years, numerous research communities, institutes, departments, journals, book series, and regular conferences were established. The Association of Internet Researchers (AoIR) was founded in 1999 and currently its mailing list has more than 5,000 subscribers. Beside its regular conferences, the activity of the International Association for Computing and Philosophy (IACAP), the meetings of the ICTs and Society Network, and the Conference series on Cultural Attitudes towards Technology and Communication (CATaC) can be considered as popular research platforms on the topic.
5. Within the framework of a social constructivist view on technology, this is the obvious reason that the internet is imbued with and many aspects of its nature determined by postmodern values. Ropolyi *Internet természete. Internetfilozófiai értekezés. (in Hungarian) (On the Nature of the Internet: Discourse on the Philosophy of the Internet*).
6. It is a really significant circumstance that such outstanding experts of complexity as statistical physicists or network scientists regularly contribute to the "theory" of the Internet, e.g., Barabási, *Linked: The New Science of Networks*; Barabási, *Network Science*; Pastor-Satorras and Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach*; etc.
7. Researches published on internet-related topics in the journals of traditional disciplines can be considered as typical candidates of this research category. See, e.g., Peng et al., "Mapping the Landscape of Internet Studies: Text Mining of Social Science Journal Articles 2000–2009."
8. Hunsinger, Klastrup, and Allen, *International Handbook of Internet Research*; Consalvo and Ess, *The Handbook of Internet Studies*.
9. See, e.g., Reips and Bosnjak, *Dimensions of Internet Science*.
10. Tsatsou, *Internet Studies: Past, Present and Future Directions*.
11. See Castells, *The Rise of The Network Society*; Castells, *The Internet Galaxy: Reflections on the Internet, Business, and Society*; Wellman and Haythornthweait, *The Internet in Everyday Life*; Barabási, *Linked: The New Science of Networks*; Barabási, *Network Science*; Bakardjieva, *Internet Society: The Internet in Everyday Life*; Lessig, *Code Version 2.0*; Feenberg and Friesen, *(Re)Inventing the Internet*; Fuchs, *Internet and Society: Social Theory in the Information Age*; Fuchs, *Digital Labour and Karl Marx*; *International Journal of Internet Science*, etc.
12. On this Aristotelian philosophical methodology and its relation to the Platonic one Hegel presented some important ideas in his History of Philosophy.
13. According to my experiences, the communities of the IACAP and the ICTs and Society Network are the most sensible public to the philosophical considerations.
14. Ropolyi, *Internet természete. Internetfilozófiai értekezés* (in Hungarian) *(On the Nature of the Internet: Discourse on the Philosophy of the Internet*); Ropolyi, "Shaping the Philosophy of the Internet"; Ropolyi, *Philosophy of the Internet: A Discourse on the Nature of the Internet*.
15. Halpin "Philosophical Engineering: Towards a Philosophy of the Web"; Monnin and Halpin, "Toward a Philosophy of the Web: Foundations and Open Problems"; Monnin and Halpin, "Toward a Philosophy of the Web: Foundations and Open Problems"; Halpin and Monnin, *Philosophical Engineering: Toward a Philosophy of the Web*; Floridi, *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality*; Floridi, *The Onlife Manifesto: Being Human in a Hiperconnected Era*.
16. Halpin, "Philosophical Engineering"; Halpin and Monnin, *Philosophical Engineering: Toward a Philosophy of the Web*.
17. Floridi,*The Fourth Revolution*; Floridi, *The Onlife Manifesto*.
18. Ess, *Philosophical Perspectives on Computer-Mediated Communication*.
19. Ess,*Digital Media Ethics*.
20. Dreyfus, *On the Internet*.
21. Feenberg and Friesen, *(Re)Inventing the Internet*.
22. Ropolyi, "Virtuality and Reality—Toward a Representation Ontology."
23. Notice that the collection of papers on Philoweb was first published in the journal *Metaphilosophy* 43, no. 4 (2012). These papers are practically the same ones which are included in Halpin and Monnin, *Philosophical Engineering: Toward a Philosophy of the Web*.
24. For a more detailed discussion of the philosophical issues involved, see Ropolyi, *Az Internet természete. Internetfilozófiai értekezés* (in Hungarian) or its online English translation, (Ropolyi *On the Nature of the Internet: Discourse on the Philosophy of the Internet*.
25. Ropolyi, "Virtuality and Reality."

## REFERENCES

Bakardjieva, M. *Internet Society: The Internet in Everyday Life*. London: Sage, 2005.

Barabási, A.-L. *Linked: The New Science of Networks*. Cambridge: Perseus Books, 2002.

———. *Network Science*. Cambridge: Cambridge University Press, 2016. http://barabasi.com/networksciencebook/.

Bijker, W. E., T. P. Hughes, and T. Pinch. *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. Cambridge, MA: The MIT Press, 1987.

Castells, M. *The Rise of The Network Society*, 2nd ed. Oxford: Blackwell, 2000.

———. *The Internet Galaxy: Reflections on the Internet, Business, and Society*. New York: Oxford University Press, 2001.

Consalvo, M., and Ch. Ess. *The Handbook of Internet Studies*. Malden/Oxford/Chicester: Wiley Blackwell, 2013.

Dreyfus, H. *On the Internet*, 2nd ed. London New York: Routledge, 2009.

Ess, C. *Philosophical Perspectives on Computer-Mediated Communication*. Albany: State University of New York Press, 1996.

———. *Digital Media Ethics*. Revised and updated 2nd ed. Cambridge, Malden, MA: Polity Press, 2013.

Feenberg, A., and N. Friesen. *(Re)Inventing the Internet: Critical Case Studies*. Rotterdam: Sense Publishers, 2011.

Floridi, L. *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality*. Oxford: Oxford University Press, 2014.

———. *The Onlife Manifesto: Being Human in a Hiperconnected Era*. New York: Springer, 2015.

Fuchs, C. *Internet and Society: Social Theory in the Information Age*. London New York: Routledge, 2008.

———. *Digital Labour and Karl Marx*. New York: Routledge, 2014.

Halpin, H. "Philosophical Engineering: Towards a Philosophy of the Web." *APA Newsletter on Philosophy and Computers* 7, no. 2 (2008): 5–11.

Halpin, H., and A. Monnin. *Philosophical Engineering: Toward a Philosophy of the Web*. Chichester/Malden/Oxford: Wiley Blackwell, 2014.

Hunsinger, J., L. Klastrup, and M. Allen. *International Handbook of Internet Research*. Dordrecht: Springer, 2010.

Lamb, R., and R. Kling. "Reconceptualizing Users as Social Actors in Information Systems Research." *MIS Quarterly* 27, no. 2 (2003): 197–236.

Lessig, L. *Code Version 2.0*. New York: Basic Books, 2006.

Monnin, A., and H. Halpin. "Toward a Philosophy of the Web: Foundations and Open Problems." *Metaphilosophy* 43, no. 4 (2012): 361–79.

———. "Toward a Philosophy of the Web: Foundations and Open Problems." In *Philosophical Engineering. Toward a Philosophy of the Web*, 1–20. Chichester/Malden/Oxford: Wiley Blackwell, 2014.

Oudshoorn, N., and T. Pinch. *How Users Matter. The Co-Construction of Users and Technologies*. Cambridge, MA; London: The MIT Press, 2003.

Pastor-Satorras, R., and A. Vespignani. *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge: Cambridge University Press, 2004.

Peng, T. Q., L. Zhang, Z. J. Zhong, and J. J. H. Zhu. "Mapping the Landscape of Internet Studies: Text Mining of Social Science Journal Articles 2000–2009." *New Media and Society* 15, no. 5 (2012: 644–64.

Pinch, T. J., and W. E. Bijker. "The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other." *Social Studies of Science* 14, no. 3 (1984): 399–441.

Reips, U-D., and M. Bosnjak. *Dimensions of Internet Science*. Lengerich: Pabst Science Publisher, 2001.

Ropolyi, L. *Az Internet természete. Internetfilozófiai értekezés. (in Hungarian) (On the Nature of the Internet: Discourse on the Philosophy of the Internet*). Budapest: Typotex, 2006.

———. "Shaping the Philosophy of the Internet." In *Philosophy Bridging Civilizations and Cultures*, edited by S. Kaneva, 329–34. Sofia: IPhR—BAS, 2007.

———. *Philosophy of the Internet: A Discourse on the Nature of the Internet*. Budapest: Eötvös Loránd University, 2013. https://www.tankonyvtar.hu/en/tartalom/tamop412A/2011-0073_philosophy_of_the_internet/adatok.html.

———. "Virtuality and Reality—Toward a Representation Ontology." *Philosophies* 1 (2016): 40–54.

Tsatsou, P. *Internet Studies: Past, Present and Future Directions*. Farnham: Ashgate, 2014.

Wellman, B., and C. Haythornthweait. *The Internet in Everyday Life*. Oxford: Blackwell, 2002.


## LINKS

Association of Internet Researchers (AoIR) (2018) https://aoir.org/

Conference series on Cultural Attitudes towards Technology and Communication (CATaC) (2014) http://blogs.ubc.ca/catac/about/

History of the Internet (2018) https://www.internetsociety.org/internet/history-internet/

Hobbes's Internet Timeline 25 (2018) https://www.zakon.org/robert/internet/timeline/

Living Internet (2017) https://www.livinginternet.com/

The ICTs and Society Network (2017) https://icts-and-society.net/

The International Association for Computing and Philosophy (IACAP) (2018) http://www.iacap.org/

# *Organized Complexity: Is Big History a Big Computation?*


Jean-Paul Delahaye

**CENTRE DE RECHERCHE EN INFORMATIQUE, SIGNAL ET AUTOMATIQUE, UNIVERSITÉ DE LILLE**

Clément Vidal

**CENTER LEO APOSTEL & EVOLUTION COMPLEXITY AND COGNITION, VRIJE UNIVERSITEIT BRUSSEL**


## 1. INTRODUCTION

The core concept of big history is the increase of complexity.[1] Currently, it is mainly explained and analyzed within a thermodynamic framework, with the concept of energy rate density.[2]

However, even if energy is universal, it doesn't capture informational and computational dynamics, central in biology, language, writing, culture, science, and technology. Energy is, by definition, not an informational concept. Energy can produce poor or rich interactions; it can be wasted or used with care. The production of computation by unit of energy varies sharply from device to device. For example, a compact disc player produces much less computation per unit of energy than a regular laptop. Furthermore, Moore's law shows that from computer to computer, the energy use per computation decreases quickly with each new generation of microprocessor.

Since the emergence of life, living systems have evolved memory mechanisms (RNA, DNA, neurons, culture, technologies) storing information about complex structures. In that way, evolution needs not to start from scratch, but can build on previously memorized structures. Evolution is thus a cumulative process based on useful information, not on energy, in the sense that energy is necessary, but

not sufficient. Informational and computational metrics are needed to measure and understand such mechanisms.

We take a computational view on nature, in the tradition of digital philosophy.[3] In this framework, cosmic evolution is essentially driven by memory mechanisms that store previous computational contents, on which further complexity can be built.

We first give a short history of information theories, starting with Shannon, but focusing on algorithmic information theory, which goes much further. We then elaborate on the distinction between *random complexity*, formalized by Kolmogorov,[4] and *organized complexity*, formalized by Bennett.[5] Kolmogorov complexity (K) is a way to measure *random complexity*, or the *informational content* of a string. It is defined as the size of the shortest program producing such a string.

This tool has given rise to many applications, such as automatic classification in linguistics,[6] automatic generation of phylogenetic trees,[7] or to detect spam.[8]

Bennett's logical depth does not measure an informational content, but a *computational content*. It measures the time needed to compute a certain string S from a short program. A short program is considered as a more probable origin of S than a long program. Because of this central inclusion of time, a high (or *deep*) value in logical depth means that the object has had a rich causal history. In this sense, it can be seen as a mathematical and computational formalization of the concept of history. More broadly construed (i.e., not within the strict formal definition), we want to show that modern informational, computational, and algorithmic theories can be used as a conceptual toolbox to analyze, understand, and explore the rise of complexity in big history.

We outline a research program based on the idea that what reflects the increase of complexity in cosmic evolution is the computational content, that we propose to assimilate with logical depth, i.e., the associated mathematical concept proposed by Bennett. We discuss this idea at different levels, formally, quasi-physically and philosophically. We end the paper with a discussion of issues related to this research program.

## 2. A VERY SHORT HISTORY OF INFORMATION THEORIES

### *2.1 SHANNON INFORMATION THEORY*

The Shannon entropy[9] of a sequence S of *n* characters is a measure of the information content of S when we suppose that every character C has a fixed probability *pr*(C) to be in position *i* (the same for every position). That is:

$$H(S) = n[-\sum_{C} pr(C)(log_2(pr(C))]$$

If we know only this probabilistic information about S, it is not possible to compress the sequence S in another sequence of bits of length less than H(S). Actual compression algorithms applied to texts do search and use many other regularities beyond the relative frequency of letters. This is why Shannon entropy does not give the real minimal length in bits of a possible compressed version of S. This minimal length is given by the Kolmogorov complexity of S that we will now introduce.

### *2.2 ALGORITHMIC INFORMATION THEORY*

Since 1965, we've seen a renewal of informational and computational concepts, well beyond Shannon's information theory. Ray Solomonoff, Andreï Kolmogorov, Leonid Levin, Pier Martin-Löf, Gregory Chaitin, Charles Bennett are the first contributors of this new science,[10] which is based on the mathematical theory of computability born with Alan Turing in the 1930s.

The Kolmogorov complexity K(S) of a sequence of symbols S is the length of the smallest program S* written in binary code and for a universal computer that produces S. This is the absolute informational content or incompressible information content of S, or the algorithmic entropy of S.

Kolmogorov complexity is also called, interchangeably, *informational content* or *incompressible informational content* or *algorithmic entropy* or *Kolmogorov-Chaitin algorithmic complexity* or *program-size complexity*.

The invariance theorem states that K(S) does not really depend on the used programming language, provided the language is universal (capable to define every computable function).

The Kolmogorov complexity is maximal for random sequences: a random sequence cannot be compressed. This is why K(S) is sometimes called *random complexity* of S.

### *2.3 LOGICAL DEPTH: COMPUTATIONAL CONTENT*

Kolmogorov complexity is an interesting and useful concept, but it is an error to believe that it measures the value of the information contained in S. Not all information is useful: for example, the information in a sequence of heads and tails generated by throwing a coin is totally useless. Indeed, if a program needs to use a random string, another random string would also do the job, which means that the particular random string chosen is not important. Kolmogorov complexity is a useful notion for defining the absolute notion of a random sequence,[11] but it does not capture the notion of organized complexity.

Charles H. Bennett has introduced another notion, the "logical depth of S." It tries to measure the real value of the information contained in S or, as he proposed, its "computational content" (to be opposed to its "informational content"). A first attempt to formulate Bennett's idea is to say that the logical depth of S, LD(S) is the time it takes for the shortest program of S, S*, to produce S.[12]

Various arguments have been formulated that make plausible that indeed the logical depth of Bennett, LD(S), is a measure of the computational content of S, or of the quantity of non-trivial structures in S. To contrast it to "random complexity," we say that it is a measure of "organized complexity."

An important property of LD(S) is the *slow growth's law*:[13] an evolutionary system S(*t*) cannot have its logical depth LD(S(*t*)) that grows suddenly. This property (which is not true for the Kolmogorov complexity) seems to correspond to the intuitive idea that in an evolutionary process, whether it is biological, cultural, or technological, the creation of new innovative structures cannot be quick.

Variants of logical depth have been explored,[14] as well as other similar ideas, such as *sophistication*,[15] *facticity*,[16] or *effective complexity*.[17] Studies have established properties of these measures, and have discussed them.[18] Importantly, results show that these various notions are closely related.[19] In this paper we focus on logical depth, whose definition is general, simple, and easy to understand.

## 3. OUTLINE OF A RESEARCH PROGRAM

### *3.1 THREE LEVELS OF ANALYSIS*

Let us first distinguish three conceptual levels of the notion of computational content: *mathematical*, *quasi-physical*, and *philosophical*.

First, we presented the notion of computational content as the logical depth, as defined by Bennett. Other formal definitions of computational content may be possible, but this one has proven to be robust. This definition has been applied to derive a method to classify and characterize the complexity of various kinds of images.[20] More applications promise to be successful in the same way as Kolmogorov complexity proved useful.

Second, we have the quasi-physical level, linking computation theory with physics.[21] This has not yet been developed in a satisfactory manner. Maybe this would require physics to consider a fundamental notion of computation, in the same way as it integrated the notion of information (used, for example, in thermodynamics). The transfer of purely mathematical or computer science concepts into physics is a delicate step. Issues relate, for example, to the thermodynamics of computation, the granularity of computation we look at, or the design of hardware architectures actually possible physically.

The concept of thermodynamic depth introduced by Seth Lloyd and Heinz Pagels is defined as "the amount of entropy produced during a state's actual evolution."[22] It is a first attempt to translate Bennett's idea in a more physical context. However, the definition is rather imprecise and it seems not really possible to use it in practice. It is not even clear that it reflects really the most important features of the mathematical concept, since "thermodynamical depth can be very system dependant: some systems arrive at a very trivial state through much dissipation; others at very non trivial states with little dissipation."[23]

Third, the philosophical level brings the bigger picture. It captures the idea that building complexity takes time and interactions (computation time). Objects measured with a deep computational content necessarily have a rich causal history. It thus reflects a kind of historical complexity. Researchers in various fields have already recognized its use.[24]

This philosophical level may also hint at a theory of value based on computational content.[25] For example, a library has a huge computational content, because it is the result of many brains who worked to write books. Burning a library can thus be said to be unethical.

### *3.2 COMPUTER SIMULATIONS*

A major development of modern science is the use of computer simulations. Simulations are essential tools to explore dynamical and complex interactions that cannot be explored with simple equations. Since the most important and interesting scientific issues are complex, simulations will likely be used more and more systematically in science.[26]

The difficulty with simulations is often to interpret the results. We propose that Kolmogorov complexity (K) and logical depth (LD) would be valuable tools to test various hypotheses relative to the growth of complexity. Approximations of K and LD have already been applied to classify the complexity of animal behavior. These algorithmic methods do validate experimental results obtained with traditional cognitive-behavioral methods.[27]

For an application of K-complexity and LD to an artificial life simulation, see, for example, the work of Gaucherel comparing a Lamarkian algorithm with a Darwinian algorithm in an artificial life simulation. Gaucherel proposes the following three-step methodology:

> (1) identification of the shortest program able to numerically model the studied system (also called the Kolmogorov–Solomonoff complexity); (2) running the program, once if there are no stochastic components in the system, several times if stochastic components are there; and (3) computing the time needed to generate the system with LD complexity.[28]

More generally, in the domain of Artificial Life, it is fundamental to have metric monitoring if the complexity of the simulated environment really increases. Testing the logical depth of entities in virtual environments would prove very useful.

### *3.3 EMERGY AND LOGICAL DEPTH*

In systems ecology, an energetic counterpart to the notion of computational content has been proposed. It is called *emergy* (with an "m") and is defined as the value of a system, be it living, social, or technological, as measured by the solar energy that was used to make it.[29] This is very similar to the logical depth, defined by the quantity of computation that needs to be performed to make a structured object.

Does this mean that energetic content (emergy) and computational content are one and the same thing? No, and one argument amongst many others is that the energetic content to produce a computation diminishes tremendously with new generations of computers (c.f. Moore's law).

## 4. DISCUSSION

We formulate here a few questions that the reader may have, and propose some answers.

**Before the emergence of life, does cosmic evolution produces any computational content?**

Yes, but the memorization of calculus is nonexistent or very limited. A computation does not necessarily mean a computation with memorization. For example, atoms such as H or molecules such as $H_2O$ are all the same; there is no memory of what has happened to a particular atom or molecule. What lacks in these cases is computation with a memory mechanism.

The increase of complexity accelerates with the emergence of more and more sophisticated and reliable memory mechanisms. In this computational view, the main cosmic evolution threshold is the emergence of life, because it creates a memory mechanism in the universe (RNA/DNA). From a cosmic perspective, complexity transitions have decelerated from the Big Bang to the origin of life, and started to accelerate since life appeared.[30] The emergence of life thus constitutes the tipping point in the dynamics of complexity transitions.

Furthermore, evolutionary transitions are marked with progress in the machinery to manipulate information, particularly regarding the *memorization* of information.[31] For example, we can think of RNA/DNA, nervous systems, language, writing, and computers as successive revolutions in information processing.

**Why would evolution care about minimal-sized programs?**

We care about *short* programs, not necessarily minimally sized programs proven to be so. The shortest program (or a near shortest program) producing S is the most probable origin for S. Let us illustrate this point with a short story. Imagine that you walk in the forest, and find engraved on a tree trunk 1,000,000 digits of π, written in binary code. What is the most probable explanation of this phenomenon? There are $2^{1,000,000}$ strings of the same size, so the chance explanation has to be excluded. The first plausible explanation is rather that it is a hoax. Somebody computed digits of π, and engraved them here. If a human did not do it, a physical mechanism may have done it, that we can equate with a short program producing π. The likely origin of the digits of π is a short program producing them, not a long program of the kind print(S), which would have a length of about one million.

Another example from the history of science is the now refuted idea of spontaneous generation.[32] From our computational perspective, it would be extremely improbable that sophisticated and complex living systems would appear in a few days. The slow growth law says that they necessarily needed time to appear.

**Couldn't you have a short program computing for a long time, with a trivial output, which would mean that a trivial structure would have a deep logical depth?**

Of course, programs computing a long time and producing a trivial output are easy to write. For example, it is easy to write a short program, computing for a long time, and producing a sequence of 1,000 zeros. This long computation wouldn't give the logical depth the string, because there is also a shorter program computing much more rapidly and producing these 1,000 zeros. This means that objects with a deep logical depth can't be trivial.

**Why focus on decompression times and not compression times?**

The compression time is the time necessary to resolve a problem: knowing S, find the shortest (or a near shortest) program producing S.

By contrast, the decompression time is the time necessary to produce the sequence S from a near shortest program that produces S. It is thus a very different problem from compression.

If we imagine that the world contains many explicit or implicit programs—and we certainly can think of our world as a big set of programs producing objects—then the probability of an encounter with a sequence S depends only on the time necessary for a short program to produce S (at first glance, only short programs exist).

**Complexity should be defined dynamically, not statically.**

A measure is by definition something static, at one point in time. However, we can compare two points in time, and thus study the relative LD, and the dynamics of organized complexity.

Let us take a concrete example. What is the difference in LD-complexity between a living and a dead body? At the time of death, the computational content would be almost the same for both. This is because the computational content measures the causal history. A dead person still has had a complex history. Other metrics may be used to capture more dynamical aspects such as informational flows or energy flows.

## 5. CONCLUSION

To sum up, we want to emphasize again that random complexity and organized complexity are two distinct concepts. Both have strong theoretical foundations and have been applied to measure the complexity of particular strings. More generally, they can be applied in practice to assess the complexity of some computer simulations. In principle, they may thus be applied to any physical object, given that it is modeled digitally or in a computer simulation.

Applied to big history, organized complexity suggests that evolution retains computational contents via memory mechanisms, whether they are biological, cultural, or technological. Organized complexity further indicates that major evolutionary transitions are linked with the emergence of new mechanisms that compute and memorize.

Somewhat ironically, complexity measures in big history have neglected history. We have argued that the

computational content, reflecting the causal history of an object and formalized as logical depth—as defined by Bennett—is a promising complexity metric in addition to existing energetic metrics. It may well become a general measure of complexity.

## NOTES

1. D. Christian, *Maps of Time: An Introduction to Big History*.
2. E. J. Chaisson, *Cosmic Evolution: The Rise of Complexity in Nature*; E. J. Chaisson, "Energy Rate Density as a Complexity Metric and Evolutionary Driver."
3. K. Zuse, *Calculating Space*; G. J. Chaitin, *Meta Math!*; Seth Lloyd, *Programming the Universe: A Quantum Computer Scientist Takes on the Cosmos*; S. Wolfram, *A New Kind of Science*; L. Floridi, *The Blackwell Guide to the Philosophy of Computing and Information*.
4. Andrei N. Kolmogorov, "Three Approaches to the Quantitative Definition of Information."
5. C. H. Bennett, "Logical Depth and Physical Complexity."
6. R. Cilibrasi and P. M. B. Vitanyi, "Clustering by Compression"; Ming Li et al., "The Similarity Metric."
7. J. S. Varré, J. P. Delahaye, and E. Rivals, "Transformation Distances: A Family of Dissimilarity Measures Based on Movements of Segments."
8. Sihem Belabbes and Gilles Richard, "Spam Filtering without Text Analysis."
9. Claude E. Shannon, "A Mathematical Theory of Communication."
10. See Ming Li and P. M. B. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, for details.
11. Per Martin-Löf, "The Definition of Random Sequences."
12. A more detailed study and discussion about the formulation can be found in C. H. Bennett, "Logical Depth and Physical Complexity."
13. Ibid.
14. James I. Lathrop and Jack H. Lutz, "Recursive Computational Depth"; Luís Antunes, Lance Fortnow, Dieter van Melkebeek, and N. V. Vinodchandran, "Computational Depth: Concept and Applications"; David Doty and Philippe Moser, "Feasible Depth."
15. Moshe Koppel, "Complexity, Depth, and Sophistication"; Moshe Koppel and Henri Atlan, "An Almost Machine-Independent Theory of Program-Length Complexity, Sophistication, and Induction"; Luís Antunes and Lance Fortnow, "Sophistication Revisited."
16. Pieter Adriaans, "Between Order and Chaos: The Quest for Meaningful Information"; Pieter Adriaans, "Facticity as the Amount of Self-Descriptive Information in a Data Set."
17. Murray Gell-Mann and Seth Lloyd, "Information Measures, Effective Complexity, and Total Information"; Murray Gell-Mann and Seth Lloyd, "Effective Complexity."
18. Luís Antunes, Bruno Bauwens, André Souto, and Andreia Teixeira. "Sophistication vs Logical Depth"; Peter Bloem, Steven de Rooij, and Pieter Adriaans. "Two Problems for Sophistication."
19. N. Ay, M. Muller, and A. Szkola, "Effective Complexity and Its Relation to Logical Depth"; Antunes et al., "Sophistication vs Logical Depth."
20. Hector Zenil, Jean-Paul Delahaye, and Cédric Gaucherel. "Image Characterization and Classification by Physical Complexity."
21. C. H. Bennett, "What Increases When a Self-Organizing System Organizes Itself? Logical Depth to the Rescue"; Richard Phillips Feynman, *Feynman Lectures on Computation*.
22. Seth Lloyd and Heinz Pagels, "Complexity as Thermodynamic Depth."
23. C. H. Bennett, "How to Define Complexity in Physics and Why," 142.
24. Murray Gell-Mann, *The Quark and the Jaguar: Adventures in the Simple and the Complex*; Antoine Danchin, *The Delphic Boat: What Genomes Tell Us*; Melanie Mitchell, *Complexity: A Guided Tour*; John Mayfield, *The Engine of Complexity: Evolution as Computation*; Eric Charles Steinhart, *Your Digital Afterlives: Computational Theories of Life after Death*; Jean-Louis Dessalles, Cédric Gaucherel, and Pierre-Henri Gouyon, *Le Fil de La Vie: La Face Immatérielle Du Vivant*; J. P. Delahaye and C. Vidal, "Universal Ethics: Organized Complexity as an Intrinsic Value."
25. Steinhart, *Your Digital Afterlives*, chapter 73.
26. C. Vidal, "The Future of Scientific Simulations: From Artificial Life to Artificial Cosmogenesis."
27. Hector Zenil, James A. R. Marshall, and Jesper Tegnér, "Approximations of Algorithmic and Structural Complexity Validate Cognitive-Behavioural Experimental Results."
28. Cédric Gaucherel, "Ecosystem Complexity Through the Lens of Logical Depth: Capturing Ecosystem Individuality."
29. E.g., Howard T. Odum, *Environment, Power, and Society for the Twenty-First Century: The Hierarchy of Energy*.
30. Robert Aunger, "Major Transitions in 'Big' History."
31. Richard Dawkins, *River Out of Eden: A Darwinian View of Life*.
32. James Edgar Strick, *Sparks of Life: Darwinism and the Victorian Debates over Spontaneous Generation*.

## REFERENCES


Adriaans, Pieter. "Between Order and Chaos: The Quest for Meaningful Information." *Theory of Computing Systems* 45, no. 4 (2009): 650–74. doi:10.1007/s00224-009-9173-y.

———. "Facticity as the Amount of Self-Descriptive Information in a Data Set." *arXiv:1203.2245 [cs, Math]*, March 2012. http://arxiv.org/abs/1203.2245.

Antunes, Luís, Bruno Bauwens, André Souto, and Andreia Teixeira. "Sophistication vs Logical Depth." *Theory of Computing Systems* (March 2016): 1–19. doi:10.1007/s00224-016-9672-6.

Antunes, Luís, and Lance Fortnow. "Sophistication Revisited." In *Automata, Languages and Programming*, edited by Jos C. M. Baeten, Jan Karel Lenstra, Joachim Parrow, and Gerhard J. Woeginger, 267–77. Berlin; New York: Springer, 2003.

Antunes, Luís, Lance Fortnow, Dieter van Melkebeek, and N. V. Vinodchandran. "Computational Depth: Concept and Applications." *Theoretical Computer Science*, Foundations of Computation Theory (FCT 2003), 354, no. 3 (2006): 391–404. doi:10.1016/j.tcs.2005.11.033.

Antunes, Luís, Andre Souto, and Andreia Teixeira. "Robustness of Logical Depth." In *How the World Computes*, edited by S. Barry Cooper, Anuj Dawar, and Benedikt Löwe, 29–34. Berlin; New York: Springer, 2012.

Aunger, Robert. "Major Transitions in 'Big' History." *Technological Forecasting and Social Change* 74, no. 8 (2007): 1137–63. doi:10.1016/j.techfore.2007.01.006.

Ay, N., M. Muller, and A. Szkola. "Effective Complexity and Its Relation to Logical Depth." *IEEE Transactions on Information Theory* 56, no. 9 (2010): 4593–4607. doi:10.1109/TIT.2010.2053892. http://arxiv.org/abs/0810.5663.

Belabbes, Sihem, and Gilles Richard. "Spam Filtering without Text Analysis." In *Global E-Security*, edited by Hamid Jahankhani, Kenneth Revett, and Dominic Palmer-Brown, 144–52. Berlin; New York: Springer, 2008.

Bennett, C. H. "Logical Depth and Physical Complexity." In *The Universal Turing Machine: A Half-Century Survey*, edited by R. Herken, 227–57. Oxford University Press, 1988. https://pdfs.semanticscholar.org/ac97/5f088cf61c09bae8506808468a08467d55e6.pdf.

———. "How to Define Complexity in Physics and Why." In *Complexity, Entropy, and the Physics of Information*, edited by Wojciech H. Zurek, 137–48. Redwood City, CA: Addison-Wesley Publishing Company, 1990.

———. "What Increases When a Self-Organizing System Organizes Itself? Logical Depth to the Rescue." *The Quantum Pontiff*, February 24, 2012. http://dabacon.org/pontiff/?p=5912.

Bloem, Peter, Steven de Rooij, and Pieter Adriaans. "Two Problems for Sophistication." In *Algorithmic Learning Theory*, edited by Kamalika Chaudhuri, Claudio Gentile, and Sandra Zilles, 379–94. Springer International Publishing, 2015.

Chaisson, E. J. *Cosmic Evolution: The Rise of Complexity in Nature*. Harvard University Press, 2001.

———. "Energy Rate Density as a Complexity Metric and Evolutionary Driver." *Complexity* 16, no. 3 (2011): 27–40. doi:10.1002/cplx.20323. http://www.tufts.edu/as/wright_center/eric/reprints/EnergyRateDensity_I_FINAL_2011.pdf.

Chaitin, G. J. *Meta Math!* Atlantic Books, 2006.

Christian, D. *Maps of Time: An Introduction to Big History*. University of California Press, 2004.

Cilibrasi, R., and P. M. B. Vitanyi. "Clustering by Compression." *IEEE Transactions on Information Theory* 51, no. 4 (2005): 1523–45. doi:10.1109/TIT.2005.844059. http://arxiv.org/abs/cs/0312044.

Danchin, Antoine. *The Delphic Boat: What Genomes Tell Us*. Translated by Alison Quayle. Cambridge, MA: Harvard University Press, 2003.

Dawkins, Richard. *River Out of Eden: A Darwinian View of Life*. Basic Books, 1995.

Delahaye, J. P., and C. Vidal. "Universal Ethics: Organized Complexity as an Intrinsic Value." In *Evolution, Development and Complexity: Multiscale Evolutionary Models of Complex Adaptive Systems*, edited by Georgi Yordanov Georgiev, Claudio Flores Martinez, Michael E. Price, and John M. Smart. Springer, 2018. doi:10.5281/zenodo.1172976. https://doi.org/10.5281/zenodo.1172976.

Dessalles, Jean-Louis, Cédric Gaucherel, and Pierre-Henri Gouyon. *Le Fil de La Vie: La Face Immatérielle Du Vivant*. Paris: Odile Jacob, 2016.

Doty, David, and Philippe Moser. "Feasible Depth." In *Computation and Logic in the Real World*, edited by S. Barry Cooper, Benedikt Löwe, and Andrea Sorbi, 228–37. Berlin; New York: Springer, 2007.

Feynman, Richard Phillips. *Feynman Lectures on Computation*, edited by J. G. Hey and Robin W. Allen. Addison-Wesley Longman Publishing Co., Inc., 1998.

Floridi, L., ed. *The Blackwell Guide to the Philosophy of Computing and Information*. Blackwell Publishing, 2003.

Gaucherel, Cédric. "Ecosystem Complexity Through the Lens of Logical Depth: Capturing Ecosystem Individuality." *Biological Theory* 9, no. 4 (2014): 440–51. doi:10.1007/s13752-014-0162-2.

Gell-Mann, Murray. *The Quark and the Jaguar: Adventures in the Simple and the Complex*. New York: Freeman, 1994.

Gell-Mann, Murray, and Seth Lloyd. "Information Measures, Effective Complexity, and Total Information." *Complexity* 2, no. 1 (1996): 44–52. doi:10.1002/(SICI)1099-0526(199609/10)2:1<44::AID-CPLX10>3.0.CO;2-X.

———. "Effective Complexity." In *Nonextensive entropy–Interdisciplinary Applications*, edited by Constantino Tsallis and Murray Gell-Mann, 387–98. Oxford, UK: Oxford University Press, 2004.

Kolmogorov, Andrei N. "Three Approaches to the Quantitative Definition of Information." *Problems of Information Transmission* 1, no. 1 (1965): 1–7. doi:10.1080/00207166808803030. http://alexander.shen.free.fr/library/Kolmogorov65_Three-Approaches-to-Information.pdf.

Koppel, Moshe. "Complexity, Depth, and Sophistication." *Complex Systems* 1, no. 6 (1987): 1087–91. http://www.complex-systems.com/pdf/01-6-4.pdf.

———. "Structure." In *The Universal Turing Machine: A Half-Century Survey*, edited by Rolf Herken, 2nd ed, 403–19. New York: Springer-Verlag, 1995.

Koppel, Moshe, and Henri Atlan. "An Almost Machine-Independent Theory of Program-Length Complexity, Sophistication, and Induction." *Information Sciences* 56, no. 1 (1991): 23–33. doi:10.1016/0020-0255(91)90021-L.

Lathrop, James I., and Jack H. Lutz. "Recursive Computational Depth." *Information and Computation* 153, no. 1 (1999): 139–72.

Li, Ming, Xin Chen, Xin Li, Bin Ma, and P. M. B. Vitanyi. "The Similarity Metric." *IEEE Transactions on Information Theory* 50, no. 12 (2004): 3250–64. doi:10.1109/TIT.2004.838101. http://arxiv.org/abs/cs/0111054.

Li, Ming, and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. New York: Springer, 2008.

Lloyd, Seth. *Programming the Universe: A Quantum Computer Scientist Takes on the Cosmos*. New York: Vintage Books, 2005.

Lloyd, Seth, and Heinz Pagels. "Complexity as Thermodynamic Depth." *Annals of Physics* 188, no. 1 (1988): 186–213. doi:10.1016/0003-4916(88)90094-2.

Martin-Löf, Per. "The Definition of Random Sequences." *Information and Control* 9, no. 6 (1966): 602–19. doi:10.1016/S0019-9958(66)80018-9.

Mayfield, John. *The Engine of Complexity: Evolution as Computation*. New York: Columbia University Press, 2013.

Mitchell, Melanie. *Complexity: A Guided Tour*. New York: Oxford University Press, 2009.

Odum, Howard T. *Environment, Power, and Society for the Twenty-First Century: The Hierarchy of Energy*. New York: Columbia University Press, 2007.

Shannon, Claude E. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (1948): 379–423, 623–56.

Steinhart, Eric Charles. *Your Digital Afterlives: Computational Theories of Life after Death*. Palgrave Macmillan, 2014.

Strick, James Edgar. *Sparks of Life: Darwinism and the Victorian Debates over Spontaneous Generation*. Cambridge, MA: Harvard University Press, 2000.

Varré, J. S., J. P. Delahaye, and E. Rivals. "Transformation Distances: A Family of Dissimilarity Measures Based on Movements of Segments." *Bioinformatics* 15, no. 3 (1999): 194–202. doi:10.1093/bioinformatics/15.3.194. http://bioinformatics.oxfordjournals.org/content/15/3/194.

Vidal, C. "The Future of Scientific Simulations: From Artificial Life to Artificial Cosmogenesis." In *Death And Anti-Death*, edited by Charles Tandy, 6: Thirty Years After Kurt Gödel (1906–1978), 285–318. Ria University Press, 2008. http://arxiv.org/abs/0803.1087.

Wolfram, S. *A New Kind of Science*. Champaign, IL: Wolfram Media Inc., 2002.

Zenil, Hector, Jean-Paul Delahaye, and Cédric Gaucherel. "Image Characterization and Classification by Physical Complexity." *Complexity* 17, no. 3 (2012): 26–42. doi:10.1002/cplx.20388. http://arxiv.org/abs/1006.0051.

Zenil, Hector, James A. R. Marshall, and Jesper Tegnér. "Approximations of Algorithmic and Structural Complexity Validate Cognitive-Behavioural Experimental Results." *arXiv:1509.06338 [cs, Math, Q-Bio]*, 2015. http://arxiv.org/abs/1509.06338.

Zuse, K. *Calculating Space*. Translated by MIT. Massachusetts Institute of Technology, Project MAC, 1970. ftp://ftp.idsia.ch/pub/juergen/zuserechnenderraum.pdf.

## CALL FOR PAPERS

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

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