Jan Romportl
Pavel Ircing
Eva Zackova
Michal Polak
Radek Schuster (eds.)

Beyond AI: Artificial Golem Intelligence

Proceedings of the International Conference
Beyond AI 2013
Pilsen, Czech Republic, November 12–14, 2013

ISBN 978-80-261-0275-5

Published by University of West Bohemia

Preface

"Beyond AI: Artificial Golem Intelligence" (BAI2013) is the third conference in the *Beyond AI* series. As the AGI allusion in the conference's subtitle suggests, we want to focus on Artificial General Intelligence, maybe in a slightly heretic way. Instead of asking what methods and algorithms we should explore in order to achieve real AGI, we want to find answers to this question: "Why is AGI a holy grail of the AI field?"

Quite an ambitious goal, one might say. Perhaps yes, but in case we see that we are failing in answering this question, we can always cross out the first word and stay with "Is AGI a holy grail of the AI field?". I personally think that it is. My professional background is applied AI and I can honestly testify that it is a truly intoxicating feeling when you (with a cunning Hephaestus' smile on your lips) make a sophisticated gizmo work, indeed a gizmo full of artificial neural networks, genetic algorithms, hidden Markov models, support vector machines, probabilistic grammars and a genuine set of rules of thumb called 'expert knowledge'.

But something is still missing. You watch your gizmo do its job, metaphorically maybe similar to particular isolated and strangely twisted human cognitive functions, and somewhere deep inside you start to feel tension and compulsive urge to improve it, make it faster, more accurate, more robust, more... natural? Will you ever feel fully satisfied? Maybe someone yes, but not me. Because nothing compares to a well-fed Golem, an artificially emerging human nourished from inanimate matter by Kabbalah of modern AI, a being with the unholy capacity of stealing the uniqueness of human soul. But wait – is this capacity really unholy? Isn't it the other way around?

Lets see what the speakers and the audience of BAI2013 can say about it. We have invited a group of great keynote speakers whose talks shall initiate such discussions. Abstracts of most of their talks are printed in this volume, specifically in its first part. The rest of the volume is dedicated to the full papers of the speakers who made it through the double-blind review process of our Programme Committee – my great thanks go to all of them.

Special thanks belong to our organising team: Eva Žáčková, Pavel Ircing, Michal Polák, Radek Schuster and Tzu-Keng Fu. Moreover, Pavel Ircing, who put together these proceedings, did all the typesetting, had heavenly patience with converting MS Word submissions to LaTeX and helping LaTeX-drowning philosophers, deserves such great thanks that he shall receive them chilled, freshly draught and served by a little cute AGI, just the way it is done here in Pilsen.

Pilsen, November 2013

Jan Romportl
Organising Committee Chair
BAI 2013

Organisation

Beyond AI: Artificial Golem Intelligence (BAI 2013) is organised by Department of Interdisciplinary Activities, New Technologies Research Centre, University of West Bohemia, Pilsen, Czech Republic. It is also supported by Department of Cybernetics of the same university. The conference took place in Pilsen, on November 12–14, 2013.

Programme Committee

Jiří Beran (Psychiatry Clinic, University Hospital, Pilsen)
Tarek R. Besold (Institute of Cognitive Science, University of Osnabrück)
Jan Betka (Department of Otorhinolaryngology and Head and Neck Surgery, University Hospital Motol, Prague)
Nick Campbell (School of Linguistic, Speech and Communication Sciences, Trinity College, Dublin)
David Díaz Pardo de Vera (Signals, Systems and Radiocommunication Department, Technical University of Madrid)
Hamid Ekbia (School of Library and Information Science, Indiana University, Bloomington)
Luis Hernández Gómez (Signals, Systems and Radiocommunication Department, Technical University of Madrid)
Ivan M. Havel (Centre for Theoretical Study, Prague)
Søren Holm (School of Law, University of Manchester)
Jozef Kelemen (Institute of Computer Science, Silesian University, Opava)
Kevin LaGrandeur (New York Institute of Technology)
Vladimír Mařík (Faculty of Electrical Engineering, Czech Technical University, Prague)
Peter Mikulecký (Faculty of Informatics and Management, University of Hradec Králové)

Hugo Pinto (AI Engineers Ltda, Porto Alegre)
Josef Psutka (Faculty of Applied Sciences, University of West Bohemia, Pilsen)
Raúl Santos de la Cámara (HI Iberia R&D, Madrid)
Kevin Warwick (School of Systems Engineering, University of Reading)
Yorick Wilks (Florida Institute for Human & Machine Cognition)
Enrico Zovato (Nuance Communications, Torino)

Organising Committee

Jan Romportl
Eva Žáčková
Pavel Ircing
Radek Schuster
Michal Polák
Tzu-Keng Fu

Sponsoring

The conference is supported by the project "Interdisciplinary Partnership for Artificial Intelligence" (CZ.1.07/2.4.00/17.0055), OP Education for Competitiveness, funded by the European Social Fund in the Czech Republic and by the State Budget of the Czech Republic. Further support was kindly provided by the Urban District Pilsen 3.

MĚSTSKÝ
OBVOD
PLZEŇ 3

Keynote Talks

An Approach Towards the Embodied (Nonhuman) Mind
Rachel Armstrong (University of Greenwich)

The Construction of Light
Ron Chrisley (University of Sussex)

AI, Its Metaphors, and Their Mutations
Hamid Ekbia (Indiana University)

Beyond Artificiality, Generality and Intelligence
Ben Goertzel (AGI Society)

Artificial Responsibility
J. Storrs Hall (Independent scientist and author)

Robot: The Modern Age Golem
Jana Horakova (Masaryk University, Brno)

What do Cars Think of Trolley Problems: Ethics for Autonomous Cars
Anders Sandberg (University of Oxford)

Creatures Made of Clay
Brian Cantwell Smith (University of Toronto)

METAMERGENCE – Kinetic Intelligence and Physical Thinking Beyond the Artificial-Natural Binary
Jaime del Val (Reverso, Madrid)

Sex, Lies and Games: Turing Style
Kevin Warwick (University of Reading)

What Can We Do with Crude Matter?
Joanna Zylinska (Goldsmiths, University of London)

Table of Contents

I Keynote talk abstracts

II Regular papers

Part I

Keynote talk abstracts

Sex, Lies and Games: Turing Style

Kevin Warwick

School of Systems Engineering, University of Reading
Reading, UK
k.warwick@reading.ac.uk

An important aspect of human intelligence is our communication. Machine performance in the area of mimicking human communication, is a well studied aspect of AI, both in terms of natural language understanding and philosophy. We know what the Turing test is essentially all about. But humans are complex organisms, and this is especially apparent when we communicate. How can a machine hope to appear to encapsulate some of those complexities? Through a study of practical Turing tests, looking at actual discourses, we take a look in particular here at issues such as gender identification in conversation and the effects of lying. In doing so we investigate further the usefulness of the Turing test (originally titled Turing's Imitation Game), and consider what it actually tells us both about the machines and humans involved.

AI, its Metaphors, and their Mutations

Hamid Ekbia

Indiana University
Bloomington, IN, United States
hekbia@indiana.edu

That human cognition is inherently metaphorical is as much a fact of ordinary thinking as it is of science and engineering, where metaphors loom large in the conception, articulation, and communication of ideas, models, and theories. Artificial Intelligence, as a set of scientific and engineering practices, also thrives on metaphors, putting them to good use in a generative manner. However, there are interesting differences between this use and how metaphors have been traditionally used in other areas of science and engineering. In this talk, I explore the differences on the three dimensions of modeling, implementation, and evaluation. These have to do with the relationship, respectively, between models and mathematics, abstraction and materiality, and behavior and mechanism. I discuss the broader implications of this exploration for our understanding of the economy of metaphors as social practice.

The Construction of Light

Ron Chrisley

Centre for Research in Cognitive Science, University of Sussex
Falmer, Brighton, UK
R.L.Chrisley@sussex.ac.uk

Approaching artificial general intelligence (AGI) from the perspective of machine consciousness (MC), I will briefly address as many of the topic areas of the conference as possible within the time allotted:

- The mind is extended, but Otto's beliefs are not in his notebook; Prosthetic AI vs AGI (Nature of Intelligence)
- Scepticism about MC and the threat of atrocity (Risks and Ethical Challenges)
- Theodicy, the paradox of AI, and the *Imago Dei*; Naturalising the Spiritual (Faith in AGI)
- Narrative, dreams and MC; Herbert's *Destination Void* as research programme (Social and Cultural Discourse)
- How to fix GOFAI; the Mereological Constraint on MC (Invoking Emergence)
- Artificial creativity as embodied seeking of the subjective edge of chaos (AI and Art)

METAMERGENCE – Kinetic Intelligence and Physical Thinking Beyond the Artificial-Natural Binary

Jaime del Val

Reverso, Madrid, Spain
jaimedelval@reverso.org

Artificial and intelligence are two contested terms. On the one hand the division between artificial and natural is part of a colonial history of boundary production that has privileged certain kinds of humans, on the other intelligence has too often been narrowly identified with certain rational and linear processes. But the boundary between natural and artificial is uncertain, and thinking is always an embodied non-linear kinetic process that happens across all nature-cultural strata in different forms of reflexive or non-reflexive modalities. The project METABODY will outline an inventive horizon for new forms of embodied non-linear intelligence/thinking across the nature-cultural spectrum, where rather than insisting on a model for prediction, simulation and control, we will investigate on the potential microdeviations of thinking/moving towards unpredictable potentials.

Robot: The Modern Age Golem

Jana Horakova

Faculty of Arts, Masaryk University
Brno, Czech Republic
horakova@phil.muni.cz

“Who were you if you admired him?” asked Robert Wechsler in the article Čapek in America (1992). “Well, that question was answered very soon ···You were a sci-fi fan.”

Karel Čapek’s work is considered an important part of the science-fiction history, if for no other reason than that he had invented one of the principal science fiction character, the robot, the artificial man, serially produced by the utopian factory known as the R.U.R., Rossum’s Universal Robots (1920/21).

The Robot, written with capital “R” by Karel Čapek, belongs to the family of artificial creatures accompanying man for thousand years already. We can divide the history of the technology-based representation of human into four stages. A mythic Golemic age, the age of clocks, the age of steam, and finally, the age of communication and control (Weiner, 1948). We used to connect the figure of robot, as well as a cyborg, with the latest stage of the technological progress, or even with its future stages. However, the Robot character reflects the deepest stages of the artificial man myth at the same time.

Karel Čapek was aware of it. In Prague newspaper Prager Tagblatt (September 23, 1935), he wrote, “R.U.R. is in fact a transformation of the Golem legend into a modern form. However, I realized this only when the piece was done. ‘To hell, it is the Golem in a fact,’ I said to myself. ‘Robots are factory mass produced Golem.’ ” The Robot, as a drama character, reflects both the dominant metaphor of the beginning of the 20th century, the period called the Machine age, and the spirit of the time (Zeitgeist), which caused the Golem legend revival mainly due Expressionists.

Introducing different stage productions of the RUR within years 1921 and 1924, I will follow the process of the Robot transformation from the metaphor of the state of the humanity in the Machine age into the main figure, even the symbol, of the technological progress myth. I will introduce the audience with the series of significant stage productions of the R.U.R.: The point of departure will be the official first night of the play in the National Theatre in Prague (January 25, 1921). We will continue through the New York (1922), and London (1923) stage productions, which gained worldwide fame to the play. We will stop in the Berlin (1923) and Vienna (1923) to see theatre productions, significant for their stage design particularly, and the flight will lend in Paris (1924).

The figure of Robot will be treated rather as a discursive object than as an invention of the ingenious author. However, the lecture wants to be a tribute to Karel Čapek, who celebrates 75th anniversary of the death this year.

Artificial Responsibility

J. Storrs Hall

Independent scientist and author
josh@autogeny.org

When a machine does something wrong, the current standard assumes that the machine’ s designer, builder, or operator is at fault. And yet the more closely machines resemble and act like human beings, the more closely we will think of them as human or at least human-like. We have now built autonomous machines, and ones whose actions, especially in detail, are becoming impossible for the designers and builders to have predicted. At what point, if any, will the legal and moral responsibility for its actions inhere in the machine itself? When is the robot to blame?

An Approach Towards the Embodied (Nonhuman) Mind

Rachel Armstrong

School of Architecture, Design & Construction
University of Greenwich, London, UK
r.a.armstrong@greenwich.ac.uk

Contemporary Western theories of mind are representational and forged by a range of philosophies and models that are subsequently mapped onto physical changes, which take place during cognition. The relationships between these phenomena are evidenced through dynamic imaging techniques such as, EEG and PET scans. Yet the presumed direct relationship between mental model, experience and empirical evidence, as being conceptual equivalents, gives rise to experiments where notions of mind are predetermined rather than exploratory.

My research is not directed towards developing a theory of mind per se, but exists as an experimental practice situated within the field of Natural Computing, which offers an embodied approach towards developing a computational foundation for the experience of conscious awareness by using an experimental chemical system that exhibits lifelike phenomena [1]. My work examines a range of phenomena associated with the self-assembly of a particular kind of chemical computing experienced through Bütschli droplets. These simple chemistries were first described by Otto Bütschli in 1892 and spontaneously emerge when concentrated alkali solution is added to a field of olive oil [2]. The solution spreads out and breaks up into millimeter scale droplets that exhibit lifelike chemical processes that include movement, sensitivity, the production of microstructures and population scale behaviours [3], see Figure 1.

Within this range of phenomena a range of striking behaviours are observed where droplets forge dynamic relationships with their

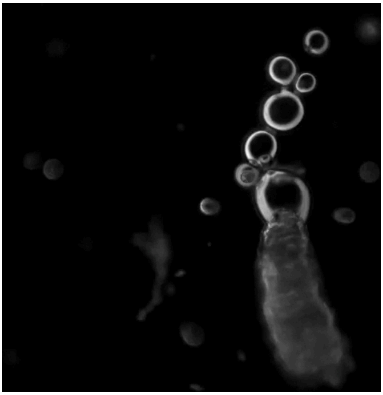

Fig. 1. Self-organizing Bütschli droplets form microstructures and exhibit lifelike behaviours such as movement, sensitivity and population scale behaviours (*Micrograph x4 magnification, Rachel Armstrong*).

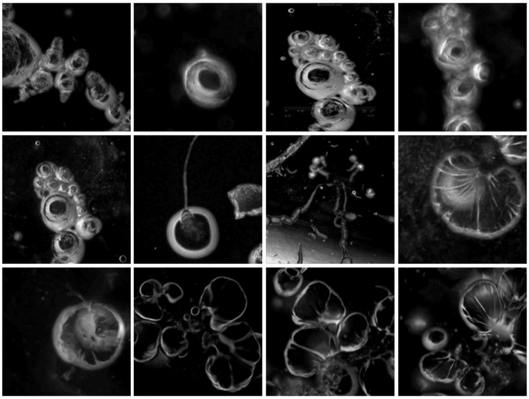

Fig. 2. Bütschli droplets self-organize into "weakly communicating" assemblages that may reach tipping points in self-organization that result in the simultaneous alteration of morphology and behavior (*Micrograph collage x4 magnification, Rachel Armstrong*).

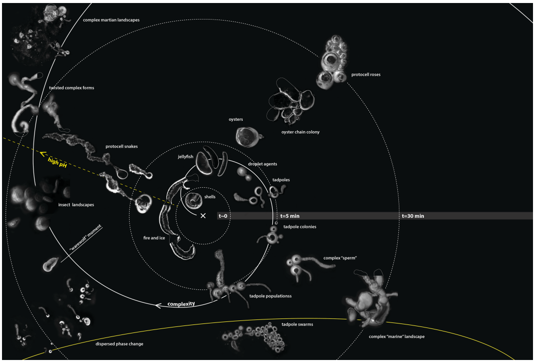

Fig. 3. The behavioural and structural groupings of Bütschli droplets are mapped according to an oceanic ontology [6] to reveal underlying patterns that may be indicative of chemical decision-making that potentially may underpin nonhuman mental processes (*Diagram by Rachel Armstrong and Simone Ferracina*).

environment and appear to be capable of "decision" making [4], see Figure 2.

The Bütschli system potentially offers an experimental model where fundamental processes that conceivably contribute to theories of mind may be explored as an extended, physical phenomenon where decision-making is dependent on environmental conditions and emergent behaviours can be directly observed and reflected back on to contemporary theories of mind. While there is no attempt to argue that these emergent phenomena constitute a "mind" in themselves, they do offer an experimental system through which the material processes that embody engaging with the self-organizing properties of chemical decision making [5]. Notably, Bütschli droplets do not attempt to represent human experience but explore the possibility of other species of "mind". Specifically then, this talk will reflect on how the observed lifelike interactions between droplet bodies may be applied as a tool for considering theories of mind that does not seek to mimic or reconstruct the human system, see Figure 3. The manner in which such embodied, chemical behaviours may exhibit recognizable and directly observable qualities associated with "mind" will also be considered.

References

1. Armstrong, R., Hanczyc, M.M.: Bütschli dynamic droplet system. Artificial Life Journal **19**(3-4) (2013) 331–346
2. Bütschli, O.: Untersuchungen über microscopische Schaume und das Protoplasma. Leipzig (1892)
3. Armstrong, R.: Unconventional computing in the built environment. International Journal of Nanotechnology and Molecular Computation **3**(1) (2011) 1–12
4. Armstrong, R.: Coming to terms with synthetic biology. Everything under control, Volume Magazine **35** (2013) 110–117
5. Adamatzky, A., Armstrong, R., Jones, J., Gunji, Y.P.: On Creativity of Slime Mould. International Journal of General Systems **42**(5) (2013) 441–457
6. Lee, M.: Oceanic ontology and problematic thought. `http://www.barnesandnoble.com/w/oceanic-ontology-and-problematic-thought-matt-lee/1105805765` (2011)

What do Cars Think of Trolley Problems: Ethics for Autonomous Cars

Anders Sandberg and Heather Bradshaw-Martin

The Future of Humanity Institute, Faculty of Philosophy
University of Oxford, Oxford, UK
anders.sandberg@philosophy.ox.ac.uk

Fully autonomous vehicles are increasingly practical, and may become common within a few decades. There are good reasons for this in terms of safety, flexibility and disability rights. However, a usable autonomous vehicle will from time to time have to make decisions that, if a human driver performed them, would have counted as moral decisions. But foreseeable vehicles will not be moral agents but rather moral proxies. So where should the moral responsibility lie? There are advantages to keeping moral responsibility as close to the user as possible.

This simplification of the chains of moral responsibility can be achieved by designing vehicles with different ethical decision profiles (within limits) and allowing users to retain part of their moral responsibility by choosing artefacts with behaviour that most closely matches their own settled convictions. This may provide a solution to artefacts which must operate in areas where humans disagree about what the best moral action is.

Part II

Regular papers

Artificial or Natural Intelligence?

Vladimír Havlík

Institute of Philosophy
The Academy of Sciences of the Czech Republic
Prague, Czech Republic
havlik@flu.cas.cz

Abstract. The distinction between natural and artificial intelligence seems to be intuitive and evident at first sight. It is not surprising that from the philosophical point of view this distinction is the basic question which fundamentally affects other considerations and conclusions connected with artificial intelligence. In this article I would like to explicate this difference and give attention to possible conclusions that result from it. I present a few examples of natural-artificial distinction in the philosophy of science and then discuss what results from it for the problem of intelligence. On the basis of Dennett's conception of intentionality I try to show that besides the traditional conception there is another perspective in which the natural-artificial distinction disappears.

Keywords: artificial intelligence, natural intelligence, artifact, natural process, intrinsic intentionality

If we are interested in the problem of artificial intelligence from the philosophical point of view we need a clear distinction between the natural and artificial in the first place. It seems to me that in many debates of artificial intelligence this distinction is not explicitly expressed and that there is only intuitive notion that something like this exists or that that the distinction is fundamental unproblematic. On the other hand I consider this difference between natural and artificial intelligence to be a basic question which fundamentally affects other considerations and conclusions of this topic. I would like to explicate this difference and give attention to possible conclusions that result from it.

How is possible to characterize the difference between natural and artificial? In the philosophy of science we can find different ways in various fields to formulate this distinction. Some of them follow a traditional conception, others try to adopt an unusual approach. I would like to give a few examples which I find instructive for this purpose, and then try to show what results from it for the problem of intelligence.

The intuitive and traditional position of the distinction between the natural and artificial can be connected with the strict separation of these opposites. The separation alone in this case is dependent on whether something is man-made or not. In one of the latest respectable textbooks of artificial intelligence we can read: “For any phenomenon, you can distinguish real versus fake, where the fake is non-real. You can also distinguish natural versus artificial. Natural means occurring in nature and artificial means made by people” [1] (p. 5). In other influential books this distinction is implicitly adopted without difficulties, e.g. Shapiro 1992 [2], Haugeland 1997 [3], Nilsson 1998 [4], Russell and Norvig 2010 [5].[1]

There are two points related to the natural-artificial distinction. The first is a question of the status of intelligence with respect to its artificiality. Is artificial intelligence only something like intelligence? In this case there is no problem in the literature with the understanding that artificial intelligence is not like “artificial flowers” [6] and that “you cannot have fake intelligence. If an agent behaves in-

[1] To be correct Haugeland is one of those authors who consider Dennett’s conception of intentionality and think that “maybe this active, rational engagement is more pertinent to whether the intentionality is original or not than is any question of natural or artificial origin” [3](p. 8). This does not mean that the natural-artificial distinction is abandoned, but it could be seen as a promising starting point.

telligently, it is intelligent” [1](p. 5). An elementary example could be more instructive. If we compare e.g. artificial flowers and natural flowers from this point of view then this distinction seems evident at first sight. Artificial flowers are not natural flowers. The artificial flowers are not living flowers; they are only imitations of living patterns. They only imitate nature. But artificial intelligence is not an imitation but intelligence in the proper sense of this word. The second question connected to the strict distinction between natural and artificial concerns the genesis, nativity or origin of intelligence. This case is more serious and there are more conceptual difficulties. Let’s continue with our elementary example.

The artificial flowers are a result of more or less sophisticated human intentional activity and skills, but for natural flowers the main effort lies in the blind powers of nature. In natural selection there is nothing intentional and this process does not follow any goal. In this sense the artificial flowers are not a living thing but only a technological artifact. And analogically we could think about all technological artifacts in this way. These artifacts “perform a practical function on the basis of a human design” [7](p. 432). From this point of view we could claim that every entity that exists as a result of human intentional activity is artificial and not natural. This is a starting point of the strict separation of natural and artificial.

However, things are not always what they seem. For instance, how natural are natural flowers? If we think about more deeply we must admit that common “natural flowers” are the result of more or less sophisticated human intentional activity and skills too. The reason lies not only in artificial selection and breeding as an intentional human activity following from the purely practical function of a man-made selection, but it can be even more artificial when considering sophisticated techniques of genetic manipulations of DNA code. In this case we have a reason to see common “natural flowers” as a technological artifact as well. That is why we cannot find anything absolutely natural in the wild:

> Over centuries and millennia of agricultural and industrial activities nature has been deeply reconfigured by humans. ‘Native forest’ in the sense of woodland absolutely untouched by human exists nowhere but in the human imagination. No part of the earth has been completely unaffected by the effects of human technologies. [8](p. 2)

On the other side technological artifacts are never really unnatural. They have a natural basis in the chemical and physical compounds or entities which compose them and as such they belong to nature:

> Technological artifacts have a dual nature. On the one hand, they are physical objects which obey the laws of nature; as physical objects their behavior can be explained in a non-teleological way. On the other hand, they are the physical embodiment of a design that has a teleological character; the whole construction is intended to perform a practical function. This function is an integral part of a technological artifact and without taking it into account, a technological artifact cannot be properly understood. [7](p. 432)

From this point of view we cannot claim that there is something which is fully unnatural: the distinction between natural and artificial is more complicated than the strict separation can encompass.

Further inspiration can be drawn from the Hacking’s conception of natural and artificial in the experimental practice of science. Opposing the traditional view, which is based on the strict separation of natural and artificial (i.e. the equipment through which nature is observed and the conditions under which it is studied are artificial whereas the objects and processes studied are natural), Hacking claims that experimentalists simply do not discover phenomena in the world but create them. “To experiment is to create, produce, refine and stabilize phenomena” [9](p. 230). From this perspective experimental phenomena are the product of intentional human activity like other technological artifacts. And if phenomena as an object of experimental research are artificial in the same way as the technological equipment through which nature is observed are, than the distinction of natural-artificial disappears. However, we must carefully formulate possible conclusions. Does Hacking's claim mean that phenomena in experiments are created like "artificial flowers"? Not in this case since the experimental phenomena are not imitations of nature but rather natural phenomena which are created under special experimental conditions. Hacking's phenomena are as common "natural flowers". They are not an imitation of nature but they are created as a result of more or less sophisticated human intentional activity and skills. That is why I cannot agree with Kroes' interpretation of Hacking's claim "to create phenomena" only in the weak sense:

> In the weak sense it means that the experimentalist creates the proper conditions for a phenomenon to take place, but does not create its specific characteristics. In the strong sense he not only causes the occurrence of the phenomenon, but also creates the specific features of the phenomenon itself. [7] (p. 435)

Kroes thinks that the weak sense of Hacking's claim could give us a way to save strict natural-artificial distinction. He says: "If we accept the weak interpretation of Hacking's expression 'creating phenomena', then we may conclude that ⋯ [phenomenon], is not created by man and therefore is not an artefact" [7](p. 435). I think that we cannot agree with this argumentation because Hacking's intervention in nature is not easily eliminable. There is not a strict border between the equipment through which nature is observed and the conditions under which it is studied and the objects which are studied. We could find many examples of this type of intervention in nature – e.g. there is an indistinguishable boundary between particle and measuring devices in quantum mechanics. Many quantum phenomena arise only under such exclusive conditions that we can hardly say that these phenomena are completely natural and that we can eliminate our part in the creation of their characteristics. My conclusion is that experimental phenomena are not natural in Kroes' weak sense interpretation without reservation, and that we must accept that the natural-artificial distinction comes in degrees. An instructive similarity can be found in the example with common flowers. The range of human intervention starts from intentional selection over several generations to the manipulation of the flowers' DNA.

As a result of foregoing considerations and in agreement with Bensaude-Vincent and Newman (2007) we must accept that "instead of opting for an absolute distinction of quality between the artificial and natural, one should accept only a gradual distinction of degree" [8](p. 2) and at the same time "the terms natural and artificial mean quite different things, depending on the context and point of view in which they occur" [8](p. 8).

However, such findings do not lead the authors (and should not lead us as well) to relativism. We can accept that the difference of natural and artificial is not as strong as we might intuitively think, that this difference is evolving and contextually dependent, but if this distinction between these opposites should have some sense for our knowledge, we must be able to determine conditions in which it is possible to identify something as a natural or artificial. Contextuality does not mean unknowability. What I see as an important finding is that "thing's 'naturalness' and 'artificiality' has one meaning when we are talking about its origin (extracted from nature versus human-made) and quite another when we are discussing its inherent qualities" [8](p. 8). These cases of confusing conclusions are not fatal when we are able to differentiate the correct context for every claim. Thus, the distinction between the natural and the artificial from this point of view is primarily a genetic one and is based on the presupposition that in some respect a human is not part of nature. We will return to this controversial claim later.

Another example which I find instructive is the connection between artificial and natural selection in Darwin's theory of evolution. Darwin in his *Origin of Species* [10] started his argument for natural selection with the description of variation of species under domestication, i.e. artificial selection. We must note that the reason Darwin initially argued on the basis of artificial selection and only later, analogically, on the basis of natural selection, was clearly strategic: Darwin was trying to gain support for his discovery of natural selection as the cause of the inception of species and evolutionary changes. Darwin traces his path to the discovery of natural selection in such a way as to lead the reader more easily to the main and controversial parts of his theory [11](p. 225). Artificial selection, e.g. the methodical accumulative selection of animals is commonly practiced and Darwin wants to show that when something as an artificial selection is possible then it is not improbable that the principle of natural selection is possible as well. He says:

> Slow though the process of selection may be, if feeble man can do much by his powers of artificial selection, I can see no limit to the amount of change, to the beauty and infinite complexity of the coadaptations between all organic beings, one with another and with their physical conditions of life, which may be effected in the long course of time by nature's power of selection. [10](p. 109)

The distinction of artificial and natural in this context lies only in man's intervention into originally spontaneous process. In this case "artificiality" again does not mean imitation. A man does not create unnatural organisms and their forms but only regulates and tries to keep required direction of changes in a fully natural process. We can assume that the same process of changes could occur without human accumulative selection under some special conditions. There is no fundamental difference in a mechanism of these two processes and actually there is only one process of selection. Moreover, we have a tendency to feel the difference that the natural is spontaneous and the artificial is affected by intentional goals. However, in this case we need to adopt another perspective which shows us unity of the selectionist process. If we accept the concept of evolution, then we do not have a problem with the principle of natural selection. The principle says that the spontaneous selectionist process affects all organic forms. Selection is a result of adaptation to the environment which includes not only the area and resources but other organisms as well. We do not have a problem imagining predators or another type of selectionist pressures from organisms which lead to the accumulation of gradual changes of some species. But why do we assume so when the source of this selectionist pressure is a man the process is artificial? It means that humans are separated from the naturalness of nature and become something that does not act in accordance with nature alone. I would like to stress that I do not want to hide the differences between human and other animals or organic forms but I think that in some cases these differences are not determinate and the right understanding is accessible only when we radically change our perspective. Then artificial selection is not a special unnatural process but only one universal process of evolution in which "Natural Selection has been the main but not exclusive means of modification" [10](p. 6).

What is important in this context for artificial intelligence? As we can see in the previous examples we pay attention to the natural and artificial distinction in the context of their origin in the first place. What could we apply in the case of intelligence? The traditional approach to this distinction is based on idea that natural intelligence is a product of biological evolution which generates organisms with intelligence[2] at some stage of the selectionist (blind) process and

[2] There is no general agreement about what is meant by intelligence, but for current needs (when we are interested in the natural-artificial distinction) we could take this as a property of organisms, e.g. it could mean that these organisms show some type of behavior. We can say very simply and more generally: for entities or actors (biological and non-biological) there is a need to interact with their surroundings on the basis of their experience and this goal leads them to the actions that we call intelligent. This is in agreement with Churchland's claim that "⋯ one obvious measure of the degree of intelligence that any creature has achieved is *how far* into the future and across *what range* of phenomena

that artificial intelligence is a man-made (intentional) project dealing with the creation of intelligence on a non-biological basis (e.g. in machines, computers and robots). To achieve this goal there is a range of possibilities: by imitation (Turing test for computers which try to act as people); by simulation or emulation (the brain can be simulated or emulated on non-biological basis); by emergence (intelligence or consciousness is obtained as an emergent property of a complex physical system); by evolution (it is possible to evolve artificial brains in machines). From the hierarchical point of view there are two basic directions: start with high-level intelligence and reproduce it directly in a computer or from bottom up, i.e. from things that are very simple pieces of intelligence to the complexity of the brain's intelligence. It is irrelevant now which mode or which combination of modes could be more promising. What we want to see from the philosophical point of view is the distinction between the natural and the artificial intelligence. As we can see in the previous examples the distinction can have a range of graduality and it could disappear in an appropriately chosen perspective.

From this point of view people strictly divide the area of natural and artificial. But their position is two-sided. On the one hand a person is a product of a natural process (selection), on the other hand a person creates culture and in some interpretation this activity separates them from nature or even puts then against nature as something that does not act in accordance with it. I think that this argumentation is false in some cases and that artificial intelligence could show us another philosophical perspective in which we could see the unity of intelligence.

This philosophical perspective could find support in Dennett's conception of intentionality [13, 14]. Dennett does not agree with Searle in the question of derived and intrinsic intentionality. Searle considers only a strict distinction between intrinsic (natural) and

it is capable of projecting the behavior of its environment, and thus how far in advance of that future it can begin to execute appropriately exploitative and manipulative motor behavior" [12](pp. 65–66).

derived (artificial) intentionality, meaning that that "the derived intentionality of our artifactual representations is parasitic on the genuine, original, intrinsic intentionality that lies behind their creation" [14](p. 50). What is attractive in Dennett's conception is the insight into the genesis of intentionality and understanding that the prime apparent distinction between intrinsic and derived forms could disappear. First, we focus on intrinsic intentionality. For Searle this is the only genuine, original intentionality (i.e. natural). But Dennett asks where genuine intentionality comes from. His answer is acceptable and refers to the creative natural process of evolution:

> [T]he brain is an artifact, and it gets whatever intentionality its parts have from their role in the ongoing economy of the larger system of which it is a part – or, in other words, from the intentions of its creator, Mother Nature (otherwise known as the process of evolution by natural selection). [14] (pp. 52–53)

Derived intentionality is analogical case. For Searle it is only parasitic on genuine intentionality (i.e. artificial). But Dennett, following the principle of unity, sees another possibility for derived intentionality. In the same way our brains (or minds) obtain intentionality from their creator (nature) and are able to delegate it to our artifacts, our artifacts (e.g. robots which might decide from its further "experience") will be able to delegate its intentionality to its artifacts.

> ⋯ it shows that derived intentionality can be derived from derived intentionality. It also shows how an illusion of intrinsic intentionality (metaphysically original intentionality) could arise. It might seem that the author of a puzzling artifact would have to have intrinsic intentionality in order to be the source of the artifact's derived intentionality, but this is not so. We can see that in this case, at least, there is no work left over for intrinsic intentionality to do. [14](p. 54)

If we are ready to accept Dennett's conception of intentionality then what does it mean for our problem of natural or artificial intelligence? Is it possible that the same is valid for intelligence? Dennett says:

> Once we adopt the engineering perspective, the central biological concept of function and the central philosophical concept of meaning can be explained and united. Since our own capacity to respond to and create meaning – our intelligence – is grounded in our status as advanced products of Darwinian processes, the distinction between real and artificial intelligence collapses. [13](p. 185)

It could seem that it is the same but Dennett continues and adds: "There are important differences, however, between the products of human engineering and the products of evolution, because of differences in the processes that create them" [13](p. 186).

If we take this note seriously then Dennett's views are inconsistent. I think that Dennett's analysis of intentionality leads us to more radical conclusion. What would it mean if we, as a result of the above-mentioned arguments, admitted that the distinction of natural and artificial intelligence is apparent only from this perspective and that in the reality there is only one evolutionary process which leads to diverse and plentiful forms which are the carriers of intelligence with varying intensity? This process started at some time during evolution on the Earth. For current needs it is not necessary to know when exactly this process started. However what is important is that "[w]e are descended from robots, and composed of robots, and all the intentionality we enjoy is derived from the more fundamental intentionality of these billions of crude intentional systems" [14](p. 55). Robots in this case are self-replicating macromolecules, as Dennett says "natural robots". If we obtained intentionality from these components without minds we can say that we obtained intelligence as well. However, in this "obtaining" a very specific mechanism of emergence of new properties is hidden. Our intelligence neither is the same as the intelligence of our macromolecules nor is created by aggregation of macromolecules' intelligence. But this is not the point. The point is that we can create descendants with intelligence not only on a biological basis. Why do we try or strive to create them? It is not only our ability – it is an ability of nature. Nature is responsible for this ability and we cannot usurp it. Nature gives us this ability and through us it can create non-biological entities with intelligence. Intelligence, like intentionality (in partial agreement with Dennett presupposition), cannot be strictly divided into natural and artificial, but should be understood as one natural process which creates intelligence with natural necessity. With some exaggeration we can see ourselves as natural tools of creation of the next intelligence history.

References

1. Poole, D., Mackworth, A.: Artificial Intelligence. Foundations of Computational Agents. Cambridge University Press (2010)
2. Shapiro, S.e.: Encyclopedia of Artificial Intelligence. Second edition, Wiley (1992)
3. Haugeland, J.: Mind Design II: Philosophy, Psychology, and Artificial Intelligence. Bradford Book (1997)
4. Nilsson, N.: Artificial Intelligence: A New Synthesis. Morgan Kaufmann (1998)
5. Russell, J., Norvig, P.: Artificial Intelligence. A Modern Approach. 3rd Edition, Prentice Hall Series in Artificial Intelligence (2010)
6. Sokolowski, R.: Natural and artificial intelligence. Daedalus **117**(1) (1988) 45–64
7. Kroes, P.: Science, technology and experiments; the natural versus the artificial. In: Proceedings of the Biennial Meeting of the Philosophy of Science Association. (1994) 431–440
8. Bensaude-Vincent, B., Newman, W.: The Artificial and the Natural. An Evolving Polarity. MIT (2007)
9. Hacking, I.: Representing and Intervening. Cambridge University Press (1983)

10. Darwin, C.: On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. John Murray (1859)
11. Ruse, M.: Charles Darwin's theory of evolution: An analysis. Journal of the History of Biology **8**(2) (1975) 219–241
12. Churchland, P.: Neurophilosophy at work. Cambridge University Press (2007)
13. Dennett, D.: Darwin's Dangerous Idea: Evolution and the Meanings of Life. Penguin Books (1996)
14. Dennett, D.: Kinds of Minds: Towards an Understanding of Consciousness. Basic Books (1996)

Syntactic Processing in the Behavior-Based Architecture of the Mind

Tzu-Wei Hung

Institute of European and American Studies, Academia Sinica, Taipei, Taiwan
htw@gate.sinica.edu.tw

Abstract. When considering the nature of human intelligence, it is held that although the behavior-based architecture of the mind specifies how cognition involving behaviors can emerge out of dynamic sensorimotor interactions, it fails to describe higher cognition involving language. I reject this view and explain from a philosophical aspect how this architecture may facilitate syntactic processing. My argumentative strategy is first to analyze why syntactic structure is not innate, inner, and universal in any classical sense, so the structure need not be pre-wired in the head; instead, it may be learned through interactions in linguistic environments. Next, I provide examples of how syntactic abstraction and reapplication are handled by a behavior-based system at the most basic level.

Keywords: syntactic structure, behavior-based architecture of the mind, extended mind, syntactic processing

1 Introduction

In this paper, I defend the view that a behavior-based architecture of the mind may describe syntactic processing – an essential aspect of human language capacity. The architecture of the mind refers to the functional mechanisms that are relatively fixed over time and indispensable to various mental activities [1, 2]. While the term "behavior-based" was originally coined by Brooks [3] to indicate a specific robotic architecture, the term is used more broadly hereafter to indicate the general idea that cognition emerges from the interactions between perception, action, and environment. Thus, a behavior-based architecture of the mind is a subpersonal mechanism explaining the personal mental activities in terms of the interaction between perception, action, and the world. This architecture stands in sharp contrast to the classical instruction-based architecture, in which the mind is modeled through a central planner sequentially processing symbolic representations according to well-defined instructions [4].

The behavior-based architecture is good at describing human cognitive skills related to behaviors, and has also been widely and successfully adopted in AI [3]. However, it is unclear how this architecture can facilitate higher cognition involving language and conceptual reasoning as it fails to handle the syntactic structure of language. The syntactic structure of a natural language refers to the conventional form determining how words are bound into sentences in that language. Influenced by Chomsky [5–9], it is held that the human mind evolves with an inner form of syntactic structure that is shared by all languages, known as universal grammar (UG). The fact that children with insufficient linguistic stimulus during language acquisition can nevertheless obtain language competence seems to indicate that this competence must base on genetically encoded principles of UG. This view joins hands with the computational theory of mind, in which the mind is regarded as an automatic device manipulating symbolic representations according to their syntactic properties [10–14]. Accordingly, the behavior-based architecture fails to satisfy the allegedly necessary conditions for linguistic processing – an inner representation that has a combinatorial structure with semantics and syntax (also known as classical representation) and a pre-wired, domain-specific processor for this classical representation are both considered necessary for language processing [15–17]. While the dominance of the above classical view has faded since the 1980s, it remains influential.

While many engineering solutions have been provided by AI researchers for designing the interface between symbols/representations and behavior-based systems [18–20], it remains unclear how human language processing can be described by such system. To show that the behavior-based system may describe syntactic processing without involving the so-called necessary conditions, Section 2 begins with analysis of why syntactic structure is not innate, inner, and universal in any classical sense. So the structure need not to be pre-wired in the mind, and is learnable through interacting with the world. Section 3 provides examples of how abstraction/reapplication of syntactic structure from an utterance can in principle be handled by a behavior-based system.

2 Characterizing Syntactic Structure

Whether syntactic processing can be adequately explained somewhat depends on whether syntactic structure is characterized rightly. If we can show that the structure is not innate, there is no need to assume a pre-wired language processor. If the structure is not inner, a behavior-based system can learn syntactic structure through interacting with the world. If it is not universal, there is no need to postulate a genetically encoded language faculty to guarantee this universality. Nonetheless, this is not an easy task as there are at least nine senses in which the structure could be innate [21], four senses in which it is universal [22], and three senses in which it is inner [23]. Thus, we must clarify in which senses the structure is not innate, not inner, and not universal.

First, syntactic structure is unlikely to be innate if we define innateness as genetic determination, as developmental invariance, and as not learned. Christiansen and Chater [24–26] reject the thesis that UG is decided by genetics, not learned, and insensitive to environment. According to them, a stable linguistic environment must be present for genes to adapt structurally. But the rapid change of language over time and across cultures and geographies does not provide the required stability. Christiansen et al. [27] conducted a number of computer simulations showing that language alteration occurs too fast to be genetically coded. It demonstrates that bias genes for specific structures can only be selected for quickly when language does not change. As no competing genetic bias will come to dominate when the pace of linguistic change and genetic mutation are equal, syntactic structure is unlikely to be genetically determined; but can, however, be altered by environment. In machine learning, syntactic rules can be acquired through probabilistic or statistical methods with minimal learning bias [28–32]. Thus, syntactic structure is learnable, sensitive to environment, and not determined by genes.

Secondly, syntactic structure needs not be inner in some classical sense. According to Wheeler [23], an inner language may refer to (i) a set of internal entities that represent external symbols and structures, which need to be handled by a domain-specific processing mechanism; (ii) mental rehearsal processes that are formulated in the form of language, which require neither an additional specialized mechanism nor the structure that is copied from external language; (iii) internal representations that can reproduce the syntactic structure of external language (i.e., inner surrogates), which requires no language-specific processor.

Marcus [33] argues that our cognitive system cannot exploit a syntactic tree structure to deal with language because this structure requires postal-code memory (i.e., the precise memory exploited by a desktop). But human memory is contextual, associative, and decays with time. This is exemplified by the inability of the mind to utilize the structure to understand sentences such as “People people left left” and “Farmers monkeys fear slept”, even if both are well-structured and meaningful.

Marcus’ observation [33] that our memory provides an insufficient basis for using a syntactic tree structure is correct. Inner structure in the sense (i) is indeed doubtful. However, Marcus’ conclusion that the structure plays no role in language processing, namely (ii), is problematic because the structure need not to be stored in the head. The mind may utilize that structure in the local environment and only store the inner surrogate of structure in offline processing, hence reducing the workload on memory. The mind may extend its cognitive power by using a system of symbols that does not belong to the head [34, 35]. Hence, the mind need not store details of a complex structure within, but may exploit it in the sense of (iii).

Third, Evans and Levinson [22] define four senses in which a grammar can be universal: (a) absolute universal, in which a set of grammatical principles can be found in all languages unexceptionally; (b) statistical universal, in which the set of principles can be found in most human languages; (c) absolute universal with restriction, in which if a language possesses certain principles then it will possess certain others; (d) statistical universal with restriction, in which if a language possesses certain principles then it has high probability to have certain others.

Evans and Levinson [22] argue that language is not universal in senses (a) and (b), which are presumed by most advocates of UG. For example, many languages have no adjectives, like Lao languages [36], and languages that exhibit no adverb, like Wambon of Papua, Cayuga and Tuscarora of Northern America, and Nkore-Kiga of Southern Uganda [37]. Additionally, the noun-verb distinction is not found in Salishan languages, and words in the single open category behave like predicates [38].

Evans and Levinson [22] also contend that statistic universals like (b) and (d) are weak, not only because not all 7 000 natural languages have been covered, but because it is not easy to rule out experimental noise, such as that arising from the same language family, language area, and socio-cultural factors. Even some properties that are likely to be statistically common to many languages – like recursion – are nothing but a “stable engineering solution satisfying multiple design constraints” [22]. Thus, there is no need to postulate a genetically encoded structure to guarantee universality in (a) and (c) senses.

3 Syntactic Processing in a Behavior-Based System

We have seen that the syntactic structure need not to be stored in our cognitive system, and can be learned through interacting with the world, but it does not necessarily follow that the behavior-based system can handle syntactic processing. So, what justifies our claim?

A prima facie thought is that action understanding and sentence understanding both involve segmenting constituents from continuous input flow and abstracting sequential order of these constituents. If a behavior-based system can manage the former, it should be able to handle the latter. Wolpert et al. [39] propose a motor selection mechanism (HMOSAIC) to choose the optimal motor command/ prediction in a given context. When receiving optic input from an observed action, the HMOSAIC can segment elemental movements from that action. It will randomly initiate multiple controller and predictor pairs, but only pairs issuing command/prediction that match the input will be selected. If a prediction of local trajectory frequently matches the optic input, then this trajectory is a properly segmented element of the entire action. Haruno et al.'s [40] simulation experiment also confirmed that the HMOSAIC could be trained to acquire movement sequences and select controllers in the correct sequence in response to various stimuli. Equally, when receiving the auditory flow of an utterance, there seems to be no reason why the same mechanism cannot segment elements (e.g., words) from that utterance and acquire the sequential order of words. As word order determines whether a sequence qualifies as a sentence, it functions as syntax.

However, syntactic processing cannot be so simple because word order has more restrictions than does movement order. First, unlike action production, not all physically producible phonetic words can be bound to form a sentence (e.g., "Mary John loves" is producible but not allowable by English structural conventions). Second, syntax is not just about word order. Sentences are highly structured and exhibit linguistic constituency, enabling sentence elements (words) to form into higher-order elements (phrases and clauses). Third, only if words are sorted into different syntactic categories can the system imposes acquired syntactic rules on these words. Fourth, compared to action processing, language processing does need some sort of representation to reflect the structure of utterances in offline processing. Accordingly, it is unjustified to defend syntactic processing in a behavior-based system without clarifying its *representation, word categorizing, syntax abstraction, and constituency.*

To this end, consider a behavior-based system with the following four functional components (Figure 1, marked in italic): An *input receptor* for receiving stimuli; A *motor controller* for generating output; a *reference signal* for indicating the goal or target of the system, and a *comparator* for detecting the gap between reference signal and input, so that the motor controller can improve its output to bridge the gap. But to handle syntax, this system needs some *minimal representation (MR)*, which only maps input stimuli onto specific states and can be used to describe an information change from one state to another among components. Brooks' (1999) behavior-based robotics system also involves the transformation of information in which a number is passed from one process to another and can be interpreted. If we map this system and its state to another domain, we can define an MR in which the "numbers and topological connections between processes somehow encode" [3].

We use MRs to replace classical representations, but this alone is not enough. We also need a domain-general mechanism for generating and exploiting MRs, to replace the domain-specific manipulator of classical representation. The input receptor and reference signal both generate inputs for the comparator and together function as a MR producer; while the comparator and motor controller both require input signals to be activated and hence serve as a MR consumer. This MR producer-consumer pair, resembling Millikan's [41] cognitive processing using pushmi-pullyu representations between a producer and a consumer, embodies how MR can be circulated across various information loops between the MR producer-consumer pair.

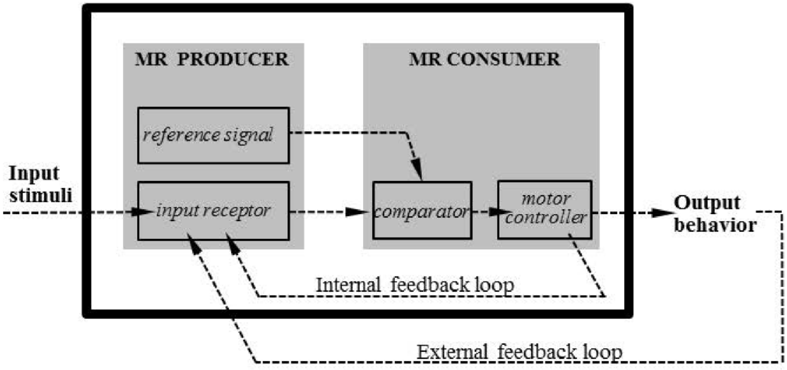

Fig. 1. The proposed behavior-based architecture of mind. Dashed lines with arrows indicate the direction of transmission routes of minimal representations.

Moreover, the system's motor controller needs to include Wolpert et al.'s [39] HMOSAIC for choosing adequate motor commands. The HMOSAIC is helpful not only in abstracting elements from continuous input flow, as we just mentioned; but also in determining what an element maps onto. Briefly, determining what a phonetic word refers to in a specific circumstance requires the system linking up the word with whatever states it is conventionally connected. To establish the link, the system's HMOSAIC may randomly initiate multiple predictions, compare these predictions with incoming input, and revise the next prediction. This processing can be repeated until the most probable referent is found. However, as the focus of the paper is not on semantics but syntax, only the latter is discussed below.

Suppose that during word segmenting and mapping, the system may detect some phonetically repeated patterns of words, which provide clues for sorting words. To decide whether words share similarities, the system must compare words according to the same criteria. These criteria amount to the system's reference signals and may partially depend on the salience of the input properties. One way to compare words is to focus on their referents (known as semantic categorization), in which words connecting to objects (i.e., nouns), motions (i.e., verbs), states of objects (i.e., adjectives), and states of motion (i.e., adverbs) can be differentiated. Another way is to focus on how repeated parts of phonetic sequences are segmented from entire sequences (known as morphological categorization). Thus, words with bound morphemes affixed to the head of word sequences (i.e., prefixes) will be separated from words with bound morphemes affixed to the end (i.e., suffixes). Moreover, words can be sorted according to their position and function in a sentence. For example, the system may detect that some words never sequentially follow certain others, and that some bind two word sequences (e.g., conjunctions and prepositions). The system's comparator will indicate whether words share similar positions and functions, and, if so, they will be grouped together. This process is known as *syntactic categorization.*

Syntactic categorization helps the system to abstract syntactic rules and to reapply acquired rules to analyze potentially infinite sentences on a finite basis. It is hardly feasible for the system using the HMOSAIC to learn and store every encountered sequence, given that the number of possible word combinations can be awfully large, let alone have the ability to output word strings according to these sequences. A promising way to solve this problem is to focus on the sequential relationships among word classes, instead of the individual words. Here, syntactic categorization can help. The system may first identify word classes (e.g., S, O, and V) in an input sequence and then learn their conventional orders, such as common word orders (e.g., either SVO or SOV) and advanced word orders (e.g., the location of a modifier in a noun phrase). The system may also determine what is conventionally forbidden, e.g., word orders that are regulated by further restrictions and exceptions. Given that conventional ordering and restrictions determine how words are combined to form sentences, they function as syntactic rules. Besides, when the system knows a rule (e.g., articles normally precede nouns if no adjectives are in-between) but is unfamiliar with an input word (e.g., "cruciverbalist" in "He is a cruciverbalist"), it can predict that the novel word is likely to be a noun than a verb and narrow down its testing of mapping. Hence, syntactic categorization and rule abstraction are mutually beneficial.

Although grasping sequential orders and restrictions allows the system to produce sentences, one may argue that syntax is not just about word order. Indeed, syntax consists of the following three elements: word order (which was just discussed), recursion (presumed by the system), and constituency. Constituency is "the bracketing of elements into higher-order elements" [22]. Thus, a noun [apple] is a constituent of a noun phrase [[the] [apple]], which is a constituent of a sentence [[John][[ate][[the][apple]]]]. How, then, can the system detect constituency? When receiving sentence S, the system may segment S into categorized elements in the order of John (N)→ → ate (V)the (D)→ apple (N) to check whether this order violates any known word orders. The system may also segment S into different constituent levels, such as N→V→NP or N→VP, and check whether S violates any conventions. On the other hand, the system may analyze the sentence in a top-down fashion. It can first segment S into two main parts, N→VP, and then segment these parts into further components, such as N→V→NP or N→V→D→N. In other words, depending on the segmenting focus, the constituency of the sentence can be detected in terms of a hierarchical analysis of sentence elements and their orders.

4 Conclusions

To summarize, in this paper I first explain why syntactic structures need not to be stored in our cognitive system and can be learned through interacting with the world. I next outline how syntactic structures of input sentences can be learned and reapplied at the most basic level. A philosophical implication of this view is that we do not need a separate mechanism for syntactic processing, and the difference between a behavior-based language processing system (e.g., human beings) and other behavior-based system (non-human animals) is not categorical but differs only in degree. Of course, more details about this behavior-based system needs to be worked out for further simulation testing, which constitutes some themes for future studies.

References

1. Byrne, M.D.: ACT-R/PM and menu selection: Applying a cognitive architecture to HCI. International Journal of Human-Computer Studies **55** (2001) 41–84
2. Pylyshyn, Z.W.: From reifying mental pictures to reifying spatial models. Behavioral and Brain Sciences **27**(4) (2004) 590–591
3. Brooks, R.: Cambrian Intelligence. MIT Press (1999)
4. Boden, M.A.: Artificial Intelligence in Psychology. MIT Press (1989)
5. Chomsky, N.: Aspects of the Theory of Syntax. MIT Press (1965)
6. Chomsky, N.: Rules and Representations. Columbia University Press (1980)
7. Chomsky, N.: Lectures on Government and Binding. Foris Publications (1981)
8. Chomsky, N.: Language from an internalist perspective. In: New Horizons in the Study of Language and Mind. Cambridge University Press (2000) 134–163
9. Chomsky, N.: Three factors in language design. Linguistic Inquiry **36** (2005) 1–22
10. Putnam, H.: Brains and behavior. American Association for the Advancement of Science, Section L (History and Philosophy of Science) (1961)
11. Fodor, J.A.: The Language of Thought. Crowell (1975)
12. Fodor, J.A.: Methodological solipsism considered as a research strategy in cognitive science. Behavioral and Brain Sciences **3** (1980) 63–73
13. Fodor, J.A.: Psychosemantics. Bradford Books (1987)

14. Fodor, J.A.: The Elm and the Expert. Bradford Books (1993)
15. Carruthers, P.: The Architecture of the Mind. Oxford University Press (2006)
16. Cosmides, L., Tooby, J.: Origins of domain-specificity. In Hirschfeld, L., Gelman, S., eds.: Mapping the Mind. Cambridge University Press (1994)
17. Pinker, S., Jackendoff, R.: The faculty of language: What's special about it? Cognition **95** (2005) 201–236
18. Hertzberg, J., Jaeger, H., Schonherr, F.: Learning to ground fact symbols in behavior-based robots. In: Proceedings of the European Conference on Artificial Intelligence, Amsterdam, Netherlands (2002) 708–712
19. MacDorman, K.F.: Grounding symbols through sensorimotor integration. Journal of the Robotics Society of Japan **17**(1) (1999) 20–24
20. Nicolescu, M.N., Mataric, M.J.: A hierarchical architecture for behavior-based robots. In: Proceedings of the First International Joint conference on Autonomous Agents and Multiagent Systems: part 1. (2002) 227–233
21. Samuels, R.: Innateness and cognitive science. Trends in Cognitive Sciences **8**(3) (2004) 136–141
22. Evans, N., Levinson, S.C.: The myth of language universals. Behavioral and Brain Sciences **32**(5) (2009) 429–492
23. Wheeler, M.: Is language the ultimate artifact? Language Science **2** (2004) 693–715
24. Chater, N., Christiansen, M.H.: Language as shaped by the brain. Behavioral and Brain Sciences **31**(5) (2008) 489–558
25. Chater, N., Christiansen, M.H.: The myth of language universals and the myth of universal grammar. Behavioral and Brain Sciences **32**(5) (2009) 452–453
26. Chater, N., Christiansen, M.H.: A solution to the logical problem of language evolution: Language as an adaptation to the human brain. In Tallerman, M., Gibson, K.R., eds.: The Oxford Handbook of Language Evolution. Oxford University Press (2012) 626–639
27. Chater, N., Christiansen, M.H., Reali, F.: The Baldwin effect works for functional, but not arbitrary, features of language. In: Proceedings of the 6th International Conference on the Evolution of Language, World Scientific Pub Co Inc, Rome (2006) 27–34

28. Parisien, C., Fazly, A., Stevenson, S.: An incremental Bayesian model for learning syntactic categories. In: Proceedings of the 12th Conference on Computational Natural Language Learning, Manchester, UK. (2008)
29. Perfors, A., Tenenbaum, J.B., Regier, T.: Poverty of the stimulus? A rational approach. In: Proceedings of the 28th Annual Conference of the Cognitive Science Society. (2006)
30. Thompson, S.P., Newport, E.L.: Statistical learning of syntax: The role of transitional probability. Language Learning and Development **3** (2007) 1–42
31. Goodman, N., Tenenbaum, J., Feldman, J., Griffiths, T.: A rational analysis of rule-based concept learning. Cognitive Science **32**(1) (2008) 108–154
32. Clark, A., Lappin, S.: Linguistic Nativism and the Poverty of the Stimulus. Oxford and Malden, MA: Wiley Blackwell (2010)
33. Marcus, G.F.: Kluge. FF (2008)
34. McClelland, J., Kawamoto, A.: Mechanisms of sentence processing: Assigning roles to constituents. In: Parallel Distributed Processing: Explorations in the Microstructures of Cognition. Volume 2., Bradford Books (1986)
35. Clark, A.: Material symbols. philosophical psychology. Philosophical Psychology **19**(3) (2006) 291–307
36. Enfield, N.J.: Adjectives. In Dixon, R.M.W., Aikhenvald, A., eds.: Adjective Classes: A Cross-Linguistic Typology. Oxford University Press (2004) 323–347
37. Hengeveld, K.: Parts of speech. In Fortescue, M., Harder, P., Kristofferson, L., eds.: Layered structure and reference in a functional perspective. (1992) 29–56
38. Jelinek, E.: Quantification in straits Salish. In: Quantification in Natural Languages. Kluwer (1995) 487–540
39. Wolpert, D., Doya, K., Kawato, M.: A unifying computational framework for motor control and social interaction. Philosophical Transactions of the Royal Society of London **B 358** (2003) 593–602
40. Haruno, M., Wolpert, D.M., Kawato, M.: Hierarchical MOSAIC for movement generation. International Congress Series **1250** (2003) 575–590
41. Millikan, R.G.: The Varieties of Meaning. MIT Press (2004)

Shared in Confidence: a Machine to a Machine (The Birth of Post-semantic Aesthetics)

Ivana Uspenski

PHD Worldwide Germany, Frankfurt, Germany
ivana.uspenski@gmail.com

Abstract. New media convergence culture has introduced significant changes to the ways in which we understand terms of reading and reception. Reading strategies are no longer products of a reader's own intentions, or the author's. Dominant readers of the new media texts based on digital code become machines. Not only do the machines read the digital texts, but they also do that based on a a digilangue, a new meta-language, which becomes a dominant language of the new software culture. The change in the very reading process demands also for the basic notion of aesthetics to be challenged. Reading digital arts means evocating a specific art experience we shall mark as procedural, post-semantic aesthetics. It rises on the horizon of a liminal transition of capitalistic culture of signification into the digital culture of software performativity, where machine intelligence takes command.

Keywords: artificial intelligence, machine, digilangue, procedural aesthetics, post-semantic aesthetics, reading, texts, digital art

1 Reading the Unreadable: Computer as the Dominant Cultural Machine

In 1997 John F. Simon Junior has started what seemed to be a relatively simple Internet based project named “Every Icon” [1]. In its essence it is a small piece of software, running on a simple algorithm. Its output represents a square, composed out of smaller squares (32 squares tall, 32 squares wide, 1 024 squares in total), of which each can be either blank or black-colored. The final aim of the algorithm is to give the output/represent all the possible combinations of either of the states for each and every of the respective 1,024 squares. However simple it sounded, and even though the algorithm itself was performing its task with a tremendous speed of 100 combinations per second, just with the elementary mathematics one was able to calculate that in order to exhaust all the combinations for only the first 32 squares, the algorithm required 16 months of continuous work. This also means, that the program Simon Jr. started almost 20 years ago, is still running with the potential to fulfill its final goal in just some 100 trillion years!

Still, the key question this simple but brilliant project has raised is that of the perception and meaning: if the algorithm is already outputting all the stages of the project, even though human brain, due to the rapidity of the process and also its virtual indefiniteness, is basically not able to perceive it, does then the process of the reading, as the perception of the given output take place at all? One can react too fast in concluding that the answer here is negative, that the perception process can take place only when an artifact is confronted with an intelligence, human intelligence. But is it really so?

Traditionally, reading is understood as one of the basic human activities which enables perception and interpretation of the artefacts of culture. Reception and consequently perception of any of the artefact of culture cannot even be imagined without its respective reading process. Therefore, it would be valid to state that one of the key attributes of any cultural artefact is that, according to the predefined set of rules, codes and conventions, it can be read and thusly included into the overall cultural discourse. This statement goes hand in hand with Escaprit's definition of reading as the process of global cultural consumption [2](p. 24). The essence of the reading process thusly understood is grounded in the affectations and meanings generation which occur once a human perceptive body is confronted with a text as a cultural artefact.

But the development and the establishment of the computer as the dominant cultural machine of today requires the traditional notion of reading to be reassessed. Even for the most elementary actions we are no longer limiting this process solely to humans. In our daily conversation we are referring to the machines reading the data, to computers processing the input. In the new media surroundings the reading and the interpretation process is no longer exclusively the heritage of a human entity, but a field in which a machine becomes an equally relevant agent. The Escarpit's operational definition of reading as cultural consumption still can be valid here, but only if we are aware of the radical change the notion of culture and cultural texts have suffered.

In his work "Code 2.0" Lawrence Lessig [3] raises a very valid example, which points in the same direction. He writes that with the ever growing misuse of the Internet and its contents, the need for the bigger control of the users accessing illegally or morally questionable content increases. But on the other hand, accessing one's personal computer is protected by the privacy laws and legislations. The question Lessig asks is then, what if authorities managed to write such an algorithm, a software able to infiltrate itself into every computer or a device upon its connection to the Internet, and which would then be able to scan all the files and folders, instantly forgetting and deleting from its own memory the benign ones, and remembering and reporting back to the authorities only the ones not in accordance with the law? The authorities, or basically people representing the authorities, as human entities would at the end get to see/interpret/read only the data referring to the illegal content, whereas the private files and folders, which the software has read, but also instantly forgotten, to them would remain inaccessible. Consequently, the real reading and interpreting agent within this process is the software itself, the machine. The very fact that the "memory" of the private files and folders was erased, does not change the fact that these were actually read. Therefore, the reading does take place even without a human agent being present or responsible for performing it.

2 Digilangue as the New Metalanguage

In order to clarify the hypothesis stated above, let us note that with the overall computation of human communication (emails replacing letters, chat-rooms and social networks replacing face-to-face talks, video-conferences replacing personal meetings, etc.) there is the overarching necessity to translate the artefacts of human culture into digital formats, in order to archive them and make them communicable, e.g. easily accessible and transferable. The digitalization of texts (a text being defined here as any cultural artefact able to undergo the process of reading and/or interpretation, albeit a poem, a sculpture or a film), as it is commonly stated, has already delivered even more revolutionary implications, than the Guttenberg invention of the printing press. Similarly as printing, digital archiving facilitated the access to otherwise very difficultly reachable content from all areas of human activity: from arts masterpieces stored in the form of a virtual, online museum, to 3D body-parts models in medicine, or even Hubble telescope photographs. And not only did the digitalization facilitate the access, it has sped it up, and even more so amplified to the unimaginable extent the processes of content sharing and distribution. In this sense, the overall digitalization one might argue has increased the efficiency and the easiness of communication.

Let us remind ourselves that the contemporary theories of communication usually refer to the mathematical definition coined by Claude Shannon, who defines communication as the processes of encoding and decoding a message which gets transmitted as a signal via a medium [4]. In this process, according to Shannon, the medium can potentially add noise to the signal carrying the message, and therefore also potentially result in misinterpretation, e.g. different understanding of the message by the receiver to that intended by the sender. Whereas Shannon has considered the communication noise as a negative element in the process, Stuart Hall emphasized a cultural definition of communication [5], where the so defined “noise” represents a positive (not in the ethical sense, but rather sense of novelty and addition) aspect of communication. He sees it as the ability of the audience to decode the message according to their own specific (predominantly social) beliefs, understandings and overall their own social context. Still, neither of the two approaches does question the very concept of the message and its uniqueness as an entity. As Manovich explains:

> Both the classical communication studies and cultural studies implicitly took for granted that the message was something complete and definite / regardless of whether it was stored in physical media (e.g. magnetic tape) or created in real time by a sender (a live TV broadcast) ⋯the “message” that the user “receives” is not just actively “constructed” by him/her (through a cognitive interpretation) but also actively managed (defining what information s/he is receiving and how). [5](pp. 33–34)

The fact is that, in order to exist and survive in the communication process of today, all messages, and consequently texts, need to be digitalized, and this means that their content needs to be decoded and encoded not in the Shannon’s sense as mediated only by the respective communication channel (e.g. telephone or a TV set), but now on an even higher level, with the help of a binary digital code, a set of ones and zeroes and additionally computed by a software, and offered via an interface to human interaction. Manovich therefore rightly expresses that we should no longer talk about the new media culture but rather about a software culture, where messages/texts get accessible only through a software, an algorithmic machine-translated function. As Manovich states:

> ···today software plays a central role in shaping both the material elements and many of the immaterial structures that together make up "culture". [5](p. 32)

He marks this shift in communication as the shift from a "message" to the one of a "platform", as the final communication experience, based on reading the text/message is not related to any single object or a text, but to the software representation of it. The text transmitted as a message does no longer have its fixed form or state, or even boundaries. Its state, form and respective experience will depend solely on the algorithm and the way the software decides to (re)present it to the user. In such a way, for example a photography stored in a folder on a computer represents only a partially arbitrary set of data, basically a row of zeroes and ones, which depending on the software used, can later on be presented not only as a photography with features different than the original, but even also as a completely different cultural form: an animation, a text or even a music composition.

"Translating" all texts into the digital code for their easier archiving and dissemination has positioned the digital code as dominant code of culture, surpassing even the up to recently unquestionably dominant position of the human language. Digital code on which all software is running becomes therefore new articulator, new codex, new linguistics, and basically new language, understood as *langue*, in the way Saussure has defined it stating that:

> If *parole* concerns the act of utterance, then *langage* concerns every conceivable *parole* generable from the system of *language* (*langue*). [6](p. 17)

This common digital language, or *digilangue* as I will be referring to it in this paper, based on its set of rules and codes, enables machines to receive an information, compute it and via an interface communicate it to a human agent, but also, which often gets forgotten, *digilangue* at the same time enables machines to easily communicate between themselves. *Digilangue* thusly becomes this primary, irreducible, communicable articulator into which, before they undergo the processes of reading and interpretation, all the texts need to be translated, including the human language (understood as *langue*) in order to make any kind of communication possible. In this sense, *digilangue* in the digital culture becomes "more mature" than the *langue*. *Digilangue* becomes the metalanguage to the human language which used to be the dominant communication meta-system. *Digilangue* overtakes the position of the key communication carrier, where the communication is understood in the Eco's words as "any flow of information from a source to a destination" [7], encompassing a role very similar to that of the *langue* as human language had in the traditional, analogue cultural systems.

The other important point is the one of the instability of the text form. As mediated by the software, the text can be accessed in variety of ways and variety of structures. The reading process is no longer necessarily linear. Linearity can only be one of its potential manifestations. The software processing the data allows jumping from one piece of data to the next, with no particular order, or arbitrary access of one data point at the time, with no predefined context. The digitalized text is read according to the actual needs and requirements, rather than according to the text's inherent predefined structure and logic. Therefore one no longer talks about reading as the cognition and cognitive interpretation process, but rather about the cultural consumption as the active management of data, as it is available and accessed at a certain point in time.

So the key question we now face is, what the implications of this newly established process are then? I dare say that they are overwhelming and that positioning the computer as the dominant cultural machine has delivered the necessary preconditions to bring the idea of dominant artificial intelligence to life. The additional argumentation behind this hypothesis is to follow below, but without the aim to put out any radical statements or wonder in the area of science fiction, the conclusion which imposes itself is that, by allowing the machines to read, we have allowed our human language, as the dominant metalanguage of culture to be read, therefore abandoning the unquestionable position of the superiority of the humankind in the cultural “food chain”. This means that humankind is no longer this one first and last step from which all the cultural communication originates and towards which it is aimed, but that a machine has taken over the role of the cultural meta-entity.

3 Machine Takes Command – the ”Phygital” World

As previously demonstrated, traditionally understood, all cultural communication was invented and intended for mankind. Even in the era of mass media, the role of the human agent in the cultural communication was central. It was the time that the theory re-addressed this concept according to the latest developments and the fact that the machines are now constituting an important part of the cultural communication flow. But what about the communication in general, how does the *digilangue* defining the communication between machines themselves?

Let us just have one example from the modern industry. One of the most mentioned concepts in the modern business of today is that of the Internet of Things. This phrase refers to the phenomenon that the overall connectivity and the Internet on a daily basis penetrate deeply into our physical reality. Within this line of thinking a concept of a “phygital” object is borne. Everyday objects like refrigerators or air conditioners, roadways or pacemakers get equipped with sensors interlinked through wired and wireless networks, often using the same protocol (IP) which connects the Internet. These objects in this way are connected to their environment, gathering the information from it and adapting their performance accordingly. The air conditioner will therefore adjust the temperature in the apartment based on the weather report it acquires from the weather Internet portal, and not according to the input given by the wo/man. Thusly, the crucial thing here is, as McKinsey puts it, that:

> ⋯these physical information systems are now beginning to be deployed, and some of them even work largely without human intervention. [8]

The “phygital” world of things around us is communicating and this communication, exchange and moreover the reception and interpretation of the information is happening without the human agent. Moreover, the human agent in this process becomes obsolete, as the machine is able to, based on the acquired set of data, draw relevant conclusions and make intelligent decisions to execute its predefined goals and tasks (like for example, perfectly adjusting the temperature in the apartment).

The question which logically follows now is whether this communication is intelligent? There is no one commonly accepted definition of intelligence that we would be able to refer to in this paper. Every definition has a specific connotation based on the context of the actual scientific discipline making it. The recent investigations in the area of artificial intelligence have brought about new interest in investigating and challenging this term. As the key aim of this paper is to demonstrate that the machines have already achieved a certain level of intelligibility which can be compared to human cognition, we will try to rely on the definitions coming both from psychology, as the discipline dealing with human mind, and the artificial intelligence scholarship.

The common trait for all psychological definitions of intelligence is that they are built within an anthropomorphic system, and they usually reflect it by referring to terms such are: life, organism, human mind, person, individual, sensory etc. This is, I believe, moreover the result of the function of the discipline, rather than the constitutive element of the definition per se. For example, Lloyd Humphreys, defines intelligence as:

> ⋯the resultant of the process of acquiring, storing in memory, retrieving, combining, comparing and using in new contexts information and conceptual skills. [9]

This definition can equally apply to machines. By reading the data, they acquire it, store it, and according to the tasks they are faced with, are able to retrieve it, combine it, even conceptualize it in order to come up with a solution to a problem.

The definitions of intelligence originating from the artificial intelligence scholarship usually refer to the same values and categories, but only instead of the notions like mind or organism, rather introduce the notions of systems, computation and algorithm. Still, the key understanding, for example for Naksashima remains very similar:

> Intelligence is the ability to process information properly in a complex environment. The criteria of properness are not predefined and hence not available beforehand. They are acquired as a result of the information processing. [9]

In both cases the intelligence is a phenomenon anchored in applying the empirical knowledge and the information historically gathered to adapt ones behavior in order make educated decisions in new situations, achieving a certain result, the success of which is measured by whether and in which level the goal was met or achieved to the extent it was planned or foreseen before taking the action. In our example of the “smart” or “intelligent” air conditioner, the measurement of the intelligence would be if and to what extent the output temperature the system is providing as a solution to the task, is congruent with what the wo/man living in the apartment would find agreeable. In this simple task, the intelligence of the machine can be compared to that of a human, by judging if the temperature set by the machine is more optimal and easier achieved than the one set manually by the human. Please note that in this paper I do not address the biological specifics of the intelligence, referring to the notions of empathy, emotions or affective traits, which still cannot be compared when it comes to human intelligence vs. the one which is being developed within machines. I will be addressing only the attributes of intelligence related to information, data and memory processing and evaluating situations in order to make concrete actions.

To come back to the hypothesis, why I do state that the first step which has made artificial intelligence even possible was allowing the machines to read? The answer to this is yet another key phenomenon of today and that is Big Data. Big Data is the term coined to describe a “collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications” [10].

The key attribute of Big Data is its vastness. As mentioned earlier in this paper, there is an overarching tendency to digitalize all the results of human activity and all cultural artefacts. This means that not only, for example, novels or paintings reproduction, but also, overall human communication activity gets mediated by the software. It also gets read by the software and moreover stored. Unlimited memory storage is additionally what makes machine intelligence superior to that of the human. And not only the unlimited memory storage, but also technically speaking, possibility not to forget, but rather keep, easily access, structure and interpret all the data ever stored. This means every time we click on a website, click on a like button or make comment on a forum our behavior pattern is memorized and stored within Big Data.

In this sense, the amount of data being uploaded via IP (Internet Protocol) is so vast and already now so historically significant that it allows the machine to simply – based on the amount of data and relevant statistical algorithms – make very reliable predictions and make not only intelligent, but also partly intelligible decisions. Just because of the very fact that the pool of data is statistically significantly more relevant that any data set a certain individual can ever have, the decisions machine make are often much more reliable and precise than any an individual might make on one own.

Now, this might just leave us at the grounds of technical expertise and precision, for which machines have already been serving mankind for more than a century. But, I argue that the implications go further than that. Just let us take one example from advertising, which has recently raised significant questions in this direction. A woman browses the Internet for certain products and certain services for couple of hours. They are predominantly food and medicals related. Suddenly, she starts getting advertisements for baby equipment and baby products, but at that point in time the woman is not aware that she is pregnant. Only two weeks later does she really finds this out. So how could a machine, specifically an advertising algorithm, be aware of this? Intuitively of course not, but based on the huge set of accumulated data, plotting the patterns of previous female consumers' online behavior, search and purchase decisions, the algorithm was able with a huge statistical relevance to predict, which other products, apart from the ones searched for, the consumer might be interested in, in near future.

In this case machine intelligence, or artificial intelligence was able to deliver even better result and more precise prediction, that the human one. Absolutely it goes without saying that this was done with no intuition, precognition or any cognitive awareness demonstrated by the advertising algorithm. The entire process was conducted without any semantic value, but as a pure procedural numerical calculation output. Therefore, yet another important differentiation needs to be introduced here – between collective intelligence (community intelligence) on one side and mass intelligence (or computable intelligence) on another.

The problematization of collective intelligence is just picking up pace, so there is no one clear insight, or one widely accepted definition of the term, one can rely on. Still, this is the term with which contemporary theory tries to describe in a more exact manner the problems of accumulation and update of human knowledge. Up to recently those were the processes which took decades, even centuries to evolve and be noticed, but which now take place in real-time, actually witnessed by the individuals and collectives who directly take part in this accumulation and update. Therefore, collective intelligence is not only a very rapid, almost instantaneous gathering and mastering of the existing information and concepts, predominantly taking place on the Internet, but rather its fundamental mark is in what Jane.McGonigal has recognized as

> ⋯work with the collected facts and viewpoints to actively author, dis-cover and invent new, computer-fuelled ways of thinking, strategizing and coordinating. [11]

It is important to note, though, that when we are talking about collective intelligences, we are not referring only to the cumulative noetic mass of individuals being part of a certain knowledge community. This term also gathers all the documents, information, conclusions and knowledge shared and stored within this community. Collective intelligence is phenomenon older than the computer culture, and can be traced far back through the history of culture, one examples amongst many being congresses or big gatherings of scientists where the latest developments and information was shared. Still, the computation and the overall connectivity the Internet has brought about, has made this process significantly faster and almost instantaneous. So on one hand it has enabled the people to get easier in touch in order to share their knowledge, developments and advance the field they are operating in further.

But, apart from being a platform facilitating human communication and information sharing, the Internet as the connected machine develops a specific intelligence system on its own. This is the term I refer to as mass-intelligence or computable intelligence. While collective intelligence is a dynamic process of human knowledge amplification, conducted by a human community online or offline with an aim to resolve a certain task, mass (computable) intelligence represents a set of computable data as gathered and stored by the machine in order to be at any given point in time fully available and susceptible to further analysis and manipulation by an algorithm aiming to resolve a task.

The key difference between collective and mass intelligence is that the latter can be categorized, quantified and all its elements can be plotted within the data base. Collective intelligence is selective. In order for it to be efficient, only the key elements and conclusions which are considered to contribute to the knowledge development the most are kept (e.g. only the top 5 papers from a conference will get published), computable mass intelligence is not. Collective intelligence is based on qualitative decision making, computable mass intelligence on quantitative parameters. And this is maybe one of the crucial differences which still makes artificial intelligence, but only on an individual level less efficient than the intelligibility of human kind: quality of communication (e.g. reliability of the data, personal experience, and intuition) of human collectives (knowledge communities) makes making extremely reliable decisions based on the relatively small set of input data possible. On an absolute level, though, simply by having access to unlimited scope of Big Data, machines in general make decisions faster. Where we stand at the moment, artificial intelligence is still on the level of computable, mass intelligence, governed by rules of calculation and quantification. The quality of the message, in the sense of its semantics and full intelligibility is yet to be achieved.

4 A Birth of a Post-semantic Aesthetics

Getting back to the example from the beginning of this paper: in the very similar manner physical objects communicate between themselves by reading and sharing the information, a machine is performing and reading all the variations of the Simon Junior's "Every Icon". As it does not have the physiological limitations of the human eye it is able to note all the outputs, no matter how fast they happen, but it is not able to recognize the patterns of meaning, like human mind does. From the net of black and white squares human mind is able to deduce, extract shapes and images, recognizing within one a specific distribution pattern, for example, a shape of a tree, a ship or even an abstract sunset. These practices account for the aesthetical activity of the human mind, where a text gets assigned a respective meaning or an affect.

Semantic-aesthetic covers the field of meaning and is the result of the process of human reading and interpretation, as they happen along the lines of human-codified quality contemplation executed through language. In our example, semantic aesthetics of the "Every Icon" project would refer to it marking the deconstruction of the expression form and questioning the limits of the new media art platforms.

On the other hand, affective aesthetics is related to the intensities, or the way cultural artefacts project their intensities and cause different bodily manifestations, resulting in sensory affectations. But can also the reading process done by a machine, quantitative in its essence, be aesthetical?

In the traditional sense of the word no, as aesthetics is an anthropocentric term, but the dawn of the artificial intelligence, I believe, requires the concept to be broadened. An aesthetic of a machine reading could potentially be defined as an intensity projected by a machine, but not sensitized, void of its human receiving agent. It cannot be affective. It can also not be contextualized or situational, or even social. It does not describe or clarify a certain artistic phenomena, nor does it analyze it within specific discursive flows. A machine aesthetic is no more than a marker of the process of intensities projection, as it is being done by a machine. Machine aesthetics is therefore a procedural aesthetics, or the aesthetics of the process.

This means that after the age of the identifying art with beauty, and then with meaning, we have come to a post-semantic era, where the aesthetics is positioned as a descriptive discipline, which marks out the inherent procedural character of an artwork, being in the constant process of re-producing its analogue self with the help of the digital code. The aesthetic object is not a text, it is a performance. And Manovich nicely differentiates it:

> I use the word performance because what we are experiencing is constructed by software in real time. So whether we are exploring a dynamic website, playing a video game, or using an app on a mobile phone to locate particular places or friends nearby, we are engaging not with predefined static documents but with the dynamic outputs of a real-time computation happening on our device and/or the server. [5]

Here Manovich is predominantly interested still in analyzing how the change is impacting human experience, but he succeeds in underlining the role of the software creating dynamic outputs, where this very process, even before it is recognized by the human, should be considered as aesthetical. “Every Icon”’s aesthetics, thusly, lays in its textualized finitely endless processuality, which exists and functions as computable flow, even without necessarily addressing a human agent.

5 Conclusion

Artificial intelligence is already amongst us. Smart phones have become our daily necessity, smart TV’s will soon be in every home, smart search allows us to in a split second get the answers to even very complicating questions, already anticipating our next move, thinking the thought even before we do. It has made the information sharing and information access faster, but also coming at a price. By delegating the biggest part of the communication tasks to machines as our extensions, we have granted them with a vast autonomy, with a space in which they freely operate on our behalf, collating the data we leave behind, patterning them, learning from them and coming up with the prediction models precise enough and good enough to allow them to make educated choices and decisions about the world which surrounds them. The air-conditioning system adjusting the temperature according to the input from a meteo report or the refrigerator putting your bottle of wine on ice, once the oven is on, are acting according to the intelligence model of making educated decisions.

One cannot deny that the education and the learning for machines is happening faster every day, and with unimaginable vaster set of data when it comes to machines than when it comes to humans. The software culture is taking over and task driven algorithms are becoming ever more reliable and correct in their outputs. Does this mean that, like in some scientific movies, we will soon need to ask ourselves, will and when human intelligence become obsolete in making the world function?

References

1. Simon, J. F.: Every icon. `http://www.numeral.com/eicon.html` (1997)
2. Escarpit, R.: Le Litteraire et le Social. ILTAM (1970)
3. Lessig, L.: Code and Other Laws of Cyberspace: Version 2.0. Basic Books (2006)
4. Chiu, E., Lin, J., McFerron, B., Petigara, N., Seshasai, S.: Mathematical theory of Claude Shannon. `http://web.mit.edu/6.933/www/Fall2001/Shannon1.pdf` (2001)
5. Manovich, L.: Sofware Takes Command: International Texts in Critical Media Aesthetics. Bloomsbury Publishing (2013)
6. Allen, G.: Intertextuality. Routledge (2000)
7. Nöth, W.: Handbook of Semiotics. Indiana Univeristy Press (1995)
8. The internet of things: McKinsey quarterly. `http://www.mckinsey.com/insights/high_tech_telecoms_internet/the_internet_of_things` (2013)
9. Legg, S., Hutter, M.: A collection of definitions of intelligence. `http://www.vetta.org/documents/A-Collection-of-Definitions-of-Intelligence.pdf` (2006)
10. Big data. `http://en.wikipedia.org/wiki/Big_data` (2006)
11. McGonigal, J.: Why I love bees: A case study in collective intelligence gaming. `http://www.avantgame.com/writings.htm` (2006)

In the Machine We Trust (Cyborg Body of Philosophy, Religion and Fiction)

Jelena Guga

Department of Interdisciplinary Activities
New Technologies Research Centre
University of West Bohemia, Pilsen
hrast78@gmail.com

Abstract. The idea of mind uploading shows that in the philosophical sense, we are still deeply embedded in Cartesian dualisms and Newtonian mechanical ways of thinking. Moreover, this idea neglects our material existence, i.e. our embodied reality, no matter how obsolete or imperfect or unable to cope with exponential technological advancements it may be. In this paper I will attempt to step out of Eurocentric and anthropocentric thought in two ways. Firstly, by introducing the Chinese philosophical concept of Tao – through the etymology of the written character Tao I will comparatively analyze it with the concept of the Machine-God. Secondly, the desire to leave the meat behind emerging from the body-mind split will be criticized through the concept of embodied consciousness. In order for a mind or any other immaterial phenomena to be uploaded into a machine, it first has to be measured, pragmatically proven and materialized. This shows the discrepancy between our mechanical hardware / the inert matter and dynamical wetware / living bodies. The paper will be an attempt to provide a platform for more inclusive, anti-essentialist ways of thinking and debating the complex and intimate relations with our machines and their potential to shape possible posthuman futures.

Keywords: Mind uploading, body-mind split, corporeality, Tao, Machine-God

The Machine feeds us and clothes us and houses us; through it we speak to one another, through it we see one another, in it we have our being.
The Machine is the friend of ideas and the enemy of superstition: the Machine is omnipotent, eternal; blessed is the Machine.
E.M. Forster: *The Machine Stops* (1909)

Judging by the present modes of technologically mediated communication we use on day-to-day basis, it is not exaggerating to say that the modern society is on a threshold of living and experiencing what Edward Morgan Forster envisioned as the future of interactions in his short story "The Machine Stops" back in 1909 [1].

It is the vision of a society so occupied and satisfied with technologically mediated communication that individuals lost the need for contact in person as well as the ability to live on the surface of the Earth, showing no intention of leaving their underground "cocoons" in which they live isolated from one another while at the same time connected through their screens, their portal to access all the information there is. The Machine contains everything and everything is perceived through the Machine. Direct experience or knowledge became unthinkable to the point of abjection and the very existence makes sense only through the Machine: "There will come a generation that had got beyond facts, beyond impressions, a generation absolutely colourless, a generation *seraphically free from taint of personality*", and whose ideas will be "far removed from that disturbing element – direct observation" [1].

Life in the technologically generated cacophony of images, sounds and textual formations characterizes a society which has overreached itself in conformism leading to increasing efficiency and decreasing intelligence. In such a setting, the word progress stands for the progress of the Machine only. Similar visions are articulated in George Orwell's novel *1984* [2] as well as in the film *Equilibrium*

(2002) for example, where human affection is seen as the source of conflicts and is therefore rendered illegal by the totalitarian system in which the only faith is the faith in the Machine (or the figure of Father in *Equilibrium* which turns out to be yet another simulation).

Although these and other similar narratives fall under the category of science fiction, they can be read as a sort of social theory of the technologically mediated world we live in today, increasingly dependent on smart phones, tablets, GPS systems, Cloud computing, Augmented Reality (AR) systems, robots, drones, nanotechnologies, etc. through which we are becoming an integral part of the omnipresence and omnipotence of the Machine. Therefore, it is not surprising that digitally coded landscapes are often seen through the prism of transcendence related beliefs and ideas as spaces of salvation from the imperfections and limits of both our material reality and our decaying mortal bodies.

In an interview for C-Theory, Paul Virilio addresses the transcendence in cyberspace as a highly complex concept [3]. While speaking in terms of metaphysics and not religion, he argues that cyberspace plays the role of God who is, and who sees and hears everything. Despite the fact that magic and religion have to a great extent been suppressed by rationality and science, one of the ironic outcomes of techno-scientific development is a renewed need for the idea of God, transcendence and salvation attributed to our machines. In other words, cyberspace and emerging technologies in general have been given the role of God or more precisely, the Machine-God: "All technologies converge toward the same spot; they all lead to a *Deus ex Machina*, a machine-God. In a way, technologies have negated the transcendental God in order to invent the machine-God." [3]

A slightly different take on cyberspace as a supreme being can be found in William Gibson's cyberpunk novel *Mona Lisa Overdrive* [4]. Gibson addresses the idea of a Godlike super-intelligence emerging in cyberspace. He writes in terms of its omniscience, omnipotence and incomprehensibility of the matrix itself, but unlike Virilio, Gibson states that the matrix is not God, "but that it has a God, since this being's omniscience and omnipotence are assumed to be limited to the matrix." [4](p. 115) Moreover, its omnipotence does not equal immortality as would ordinarily be the case in belief systems positing a supreme being because the existence of the matrix is dependent upon human agency. In that sense, it is not only that we have discovered a God-Machine, but technologically altered holy trinity: God-Human-Machine.

Naturally, the question arises whether the matrix as such could as well be regarded as intelligent or self-aware, i.e. could its processing power give rise to AGI as suggested in Gibson's novel *Neuromancer* [5] where intelligent entity Wintermute emerges from the substance and structure of cyberspace.

Would such AI be limited to digital spaces only or is it more likely that it would strongly affect our embodied existence as well, having in mind the ever increasing boundary collapse between "real" and "virtual"? To what extent would we as humans expand our ability to adapt to our environment or techno-eco systems governed by AI in this case and, more importantly, would we at all be able to tame and control it to our benefit as so often envisioned by futurist thinkers? If on top of technological cognitive and bodily human enhancements we add a techno-sentient being to the equation, how would the notion of being human transform towards the posthuman in such couplings? Finally, given that humans no longer hold the dominant central position in human-machine interfaces, would the human attributes including psyche, memory, language, cognition, etc. simply be rendered irrelevant?

Looking back at the 5000 days of existence of the World Wide Web and predicting what next 5000 days may bring, the *Wired* magazine's founding executive editor Kevin Kelly spoke of the Web as a living organism in his 2007 TED Talk [6].

Similar to Virilio's observations, he pointed out that only a few decades ago we couldn't have imagined having a direct access to any information needed, to have the whole world right before us on our digital devices. Interestingly, "it's amazing, yet we are not amazed" [6], for in such a short time we got used to the Web so quickly and effortlessly that it became almost impossible to imagine the world without access to it where we not only get the information needed but also project ourselves into it thus becoming its constitutive part on multiple levels of interaction. On top of that, we no longer rely on our own memory only but trust and rely on search engines and social networks which function as the extension of mind and self.

According to Kelly, our devices are only small windows or portals to the network as a unique machine which he refers to as the One or the One Machine. It is the most reliable machine humankind has ever built – it works continuously and its daily flow of information equals the capacity of one human brain. The only difference is that the capacity of human brain does not increase every two years, which means that in about thirty years from now, the Web will have a daily capacity of six billion human brains containing all the data we can think of. Hardware memory is already giving way to Cloud data storage accessible only via the Internet while at the same time the data location remains unknown (the users have no insight as to where exactly their data are stored nor can they be sure who has access to and control over those data). The Cloud consists of compressed abstract layers of data and is often described as the "hive-mind" emerging from hardware technical solutions and abstract software Internet models. Everything corporeal is being transcoded into the Cloud.

Consequently, we are becoming highly dependent on the abstract network of digital synapses within which everything is seemingly possible. However, we can never grasp or control it in its totality but just as with deities, we can only reflect our hopes and beliefs in its reliability. As noted by Kevin Kelly, the Web functions as synapses of the human brain and in that sense he compares it with a dynamic living organism we respond to and interact with thus giving rise to a highly complex unity in which technology acquires features of living systems:

> There is only One Machine. The Web is its OS. All screens look into the One. No bits will live outside the Web. To share

> is to Gain. Let the One read it. The One is us – we are in the One. [6]

If we use the word God here instead of the One Machine, we can clearly see the renewed need to believe in a higher power, something greater than us, i.e. the Machine-God in this case that Paul Virilio spoke of. Certainly, generalizations should be avoided when speaking of differences between various religious and/or philosophical systems which are all determined by complex social discourses throughout history, but for the purposes of this paper it will be sufficient to only outline some of the features of these systems and observe them in relation to new technologies out of which a new hybrid phenomenon is emerging, manifesting itself in the form of some sort of algorithmic religion.

It is based on our confidence in, and ever increasing dependence on the Machine or the Cloud into which we upload all of our being, knowledge, history, cultural artifacts, etc., hoping that in the very near future the emerging technologies will enable us to transgress the limitations of corporeality. It is in this sense that the Machine becomes perceived as the One, the supreme or meta-being. And just as the heavenly kingdom is a place which supports our hopes and prayers and holds a promise of eternal life, so can the Cloud be considered a place which provides a technical support for all our needs and desires which is only a click away.

Whether we are addressing monotheism, polytheism or any kind of non-institutionalized spiritual practice in the context of new technologies, digital datascapes can in a sense be considered the fields of tech-gnosis, neo-paganism and cyber-mysticism [7] where the transcendence is attainable for everyone through a screen as a portal to cyber-nirvana and back, or even to the eternal life in digital landscapes if we ever master the art of uploading the mind and leaving the materiality of our existence behind that many prophets of the future speak about.

However, the belief that technologies can save us from the sufferings of this life is not unique to the modern era and the emergence of new media technologies. Historically, every breakthrough in technological development has been up-to-date followed by predictions and visions of the future in which humanity projects itself into the seemingly limitless potentials of technology at hand, in search for the human empowerment, longevity and the essence of life. With the advent of cyberspace, many futurists were (and some still are) prone to thinking of digitally coded landscapes as “the promised land”, claiming that we will be able to leave this realm of existence and move to the place of omnipotence where the transcendence of imperfect and disappointing here-and-now awaits.

Our minds already operate with ease between the realms of imagination and reality, virtuality and materiality, but the key issue here may not be how to transcend the body for we do not know yet whether such an endeavour is at all possible, but how to incorporate the body into our “electric dreams” and bridge the gap between technological hardware and biological wetware. Regardless of the lack of scientific proof that the mind can be separated from the body to be uploaded into the machine, many futurists and especially transhumanists draw on the works of Marvin Minsky, Hans Moravec and Raymond Kurzweil and pursue the idea of mind uploading as one of the ultimate goals of technological development.

Although transhumanism is based on secular humanism and atheism, transhumanists very often use spiritual, mythological or parapsychological concepts such as the immortality of the mind/soul/spirit and transcendence of the body for instance, and attribute them to nano- and biotechnologies, robotics and information technologies, in which they see possibilities of salvation, i.e. possibility of becoming immortal through mind uploading.

Different religious and philosophical systems incorporate the idea of human immaterial substance which lives on after the death of the body. For example, Christianity negates life after death but preaches about the eternal life of the soul after death. Western thought, whether religious or philosophical, is deeply embedded in Cartesian dualism i.e. body and mind split in which the body is usually assigned negative connotations of being sinful, dirty, decaying and thus limiting to the mind, while the mind is highly idealized. On the other hand, Hinduism, Buddhism, Taoism and other religions and schools of philosophy of the East have a more holistic approach in which body and mind constitute a dynamic unity instead of opposites excluding one another.

For instance, in Chinese philosophy and religion, as well as in traditional medicine, there's a concept of Chi (氣) also known as *Prana* in Sanskrit, meaning "life force", "vitality" or "energy flow", which is immaterial or rather invisible but can at the same time be physically measured and balanced through acupuncture, breathing exercises, diet, martial arts, etc., contributing to well-being and longevity. As such, it can neither be termed immaterial or material, and it is both at the same time.

Another concept deeply embedded in Chinese philosophy and religion (Taoism, Confucianism and Zen Buddhism) is the metaphysical concept of Tao (道) which can roughly be translated as "path", "way", "principle", "logos" or "doctrine" but its essence cannot be verbalized but only experienced [8, 9]. In order to avoid a lengthy discussion on all the possible meanings of Tao, for the purpose of this paper I will only focus on the etymology of the written character Tao which shows not only the unity of body and mind but also the necessary embodied interaction with, and adaptation to environment no matter how technologically enhanced it may be.

The character Tao consists of two parts:

1. Character 首 */shou/* means "head", but if we break it further into its consisting parts (丷一自), it has a much deeper meaning. The two strokes on the top represent the principle of Yin and Yang, i.e. the interconnectedness and interdependence of the opposites (which I prefer reading as the binary code in the context of new technologies). The horizontal stroke in the middle brings these two forces together, making them complementary instead of opposing elements of a whole. The third part is character 自 */zi/* which means "self", but unlike the purely noetic nature of the self as seen and interpreted from the perspective of West-

ern thought, in Chinese cosmology the self is considered to be a unified spiritual-corporeal whole, or an embodied mind.

2. Character 辶 /*chuo*/ means to go, to walk, or to move, and it can also represent the walking surface.

Now to put the character back together, Tao as a dynamic system of Yin and Yang giving rise to each other, can be comprehended through the movement, or in other words, to become one with the Tao is to experience it through the embodied mind. This means that the isolated intellectual comprehension is not sufficient enough and it requires simultaneous bodily awareness of what is within and around us.

If we are to identify Tao with the One Machine as our omnipresent technologically mediated environment, then, comprehending the One and merging with it is not possible through uploading the mind only. For making this experience at all possible, it has to be based on corporeality, because it is through the body/embodied mind that we perceive the world and respond to it [10].

Furthermore, if the concepts of Chi, mind, soul, consciousness, etc. are placed in the context of the emerging technologies which hold the promise of transcending the body, it no longer matters whether the interpretations of the existence of our constitutive immateriality are true or false. If the mind uploading is to be made possible, we should first ask the question of where exactly the mind is located - – is it exclusively tied to the brain or is it interwoven through every cell of the body?

If the mind can exist only in relation to functional living body, does mind uploading involve hibernation or some sort of an induced coma while maintaining basic bodily functions? In a state of hibernation dominated by subconsciousness, how can the conscious processes be activated? Isn't the hardware to which the mind would eventually be uploaded also susceptible to viruses, failure and obsolescence way much faster than our imperfect, mortal bodies? Finally, in order for something to be converted and uploaded, doesn't it have to be measured, pragmatically proven or given in a sort of material form?

These and similar questions might be helpful in establishing a critical distance towards techno-enthusiasm trends often found among futurist thinkers, which seem to be a way of escapism to technological imaginary, overlooking the material aspect of corporeal existence of both our biological bodies and technological hardware which contains and generates seemingly omnipotent virtual worlds.

Likewise, this kind of attitude neglects the harsh reality of sociopolitical and economic conditions and challenges of present day. As Andy Clark has simply put it, we are by nature "products of a complex and heterogeneous developmental matrix in which culture, technology, and biology are pretty well inextricably intermingled" [11]. Embracing the materiality and virtuality of existence and stepping out of rather limiting techno-centric views might bring us closer to more profoundly revealing and fully experiencing the perpetual change in the reality of the One.

In this paper, I have introduced the notion of Tao not in terms of its religious mysticism or poetry as often understood in the West, but in its pure philosophical and linguistic form in order to initiate a discussion on the emerging technologies and potential futures not only from the perspective of linear, goal-oriented, and/or dualistic mindset, but also to include the way of thinking that is based on interconnectedness, perpetual change, networks, systems, etc.

Although originating from the traditional philosophy of Tao, these terms lay at the very foundations of the principles new emerging technologies are based on. Without any additional effort, it is possible to reconceptualize or rethink the complex man-machine relations within the context of thusly introduced concept of Tao.

The principles grounded in the Taoist philosophical thought can be offered as a sort of intuitive tool to better understand and so to govern and control the dynamics of human-machine relations. Therefore, I hope that this paper will initiate further thoughts into direction of interdisciplinary theory of technologically amplified reality based on the methodologies and terminologies merging the bodies of traditional philosophy and contemporary science.

References

1. Forster, E.M.: The machine stops. http://archive.ncsa.illinois.edu/prajlich/forster.html (1909)
2. Orwell, G.: 1984. Penguin Books (2003)
3. Wilson, L.: Cyberwar, god and television: Interview with paul virilio. http://www.ctheory.net/articles.aspx?id=62 (1994)
4. Gibson, W.: Mona Lisa Overdrive. Bantam Books (1988)
5. Gibson, W.: Neuromancer. Ace Books (1984)
6. Kelly, K.: The next 5000 days of the web. http://www.ted.com/talks/kevin_kelly_on_the_next_5_000_days_of_the_web.html (2007)
7. Cowan, D.E.: Cyberhenge: Modern Pagans on the Internet. Routledge (2005)
8. Yu-Lan, F.: A Short History of Chinese Philosophy. The Free Press (1997)
9. Needham, J.: Science and Civilization in China - History of Scientific Thought. Cambridge University Press (1956)
10. Merleau-Ponty, M.: Phenomenology of Perception. Routledge (2002)
11. Clark, A.: Natural Born Cyborgs: Minds, Technologies and the Future of Human Intelligence. Oxford University Press (2003)

The Impact of Legal Rights for AGI in the Speculative Future

Grace Halden

Birkbeck College, University of London, London, UK
gracehalden@yahoo.co.uk

Abstract. AGI is, for many, the coveted goal of the artificial intelligence field. However, once this goal is achieved a question of where the AGI fits into the human arena may be debated. One way advanced AGI may impact the human world is in regards to legal rights. My focus is not on exploring whether AGI *should* enter the courts in the quest for legal rights, but what would happen *if* this became a reality. In order to explore the theoretical concept of AGI rights, I will use science fiction and historical landmark cases to explore the issue.

Keywords: legal rights, law, science fiction, Bina48, AGI

1 Introduction

This conference asks "Why is AGI the Holy Grail of the AI field?" I reposition the question to speak about after the Grail has been won. This paper explores the legal consequences of the AGI Holy Grail. The wealth of scholarly and science fiction examples reveal a profound ambition to create an evolved intelligence and, most vitally, a will to place this AGI within the human sphere as deserving of liberty, claim, power and immunity. I also wonder if legal rights will one day be the Holy Grail *for* AGIs.

Artificial General Intelligence refers to the successful construction of intelligent machines, in which the intelligence is argued to be equal to, or surpassing, human intellect [1]. AGI is, for many, the coveted goal of the artificial intelligence field. However, once this goal is achieved a question of where the AGI fits into the human arena may be debated. One way advanced AGI may impact the human world is in regards to legal rights. Because AGI will demonstrate "general intelligence" equivalent (if not superior) to general intelligence displayed in the standard human, then the question of legal positioning may occur.

My focus is not on exploring whether AGI *should* enter the courts in the quest for legal rights, but what would happen *if* this became a reality. Thinkers such as Justin Leiber and David J. Gunkel explore whether technology can and should have rights and explore the "machine question" more broadly by looking at issues of cognition, moral agency, personhood and so on. However, my focus instead is to examine how the pursuit of rights would have an *impact* on social perspectives and on current law.

As Gunkel notes, "little or nothing has been written about the machine" in regards to rights [2]. In order for me to carve out an area of focus – that of the impact of AGI legal rights – I shall examine how science fiction foresees this issue transpiring. Science fiction has challenged the Holy Grail of AGI and has dealt extensively with the legal entanglements AGI may cause, such as: Isaac Asimov's *The Bicentennial Man* (1976) (including The *Positronic Man* by Isaac Asimov and Robert Silverberg (1992); and *The Bicentennial Man* film directed by Chris Columbus (1999)), *Star Trek: Voyager* (1995-2001), *A.I. Artificial Intelligence* (2001), and Ted Chiang's *The Lifecycle of Software Objects* (2010). These texts veer from utopic, to dystopic, to the ambiguous.

1.1 The Case of Bina48

In May 2013, the first annual conference on "Governance of Emerging Technologies: Law, Policy and Ethics" was launched which the purpose of exploring governance issues surrounding "GRINN technologies (genetics, robotics, information technology, nanotechnology, neuroscience)” [3]. The conference agenda closely concerned the variety of legal issues emerging technologies will impact. David J. Gunkel in his paper “Can a Machine Have Rights?” explored:

> Whether it is possible for a machine (defined broadly and including artifacts like software bots, algorithms, embodied robots, etc.) to have or be ascribed anything like rights, understood as the entitlements or interests of a moral subject that need to be respected and taken into account. [4]

Rather than answering this question, I instead seek to unearth the speculative problem of AGI legal rights in the current legal arena. My question is perhaps not “Can a Machine Have Rights?” but rather what will happen if such a quest is pursued.

Before Gunkel’s 2012 text asking this very question, Peter Voss spoke of the immediate importance of exploring AGI rights: “I believe that the issues surrounding the legal and moral complexity of Artificial General Intelligence are not only extremely important, but also much more urgent and imminent than many people think” [5] (p. 12). Already work is being completed on the future necessity for AGI rights. In 1985 Justin Leiber offered a dialogue entitled *Can Animals and Machines Be Persons?* During this dialogue a chimpanzee (Washoe-Delta) and a computer (Turing 346, also known as AL) fell under scrutiny and issues of personhood, moral rights and legal rights were challenged. In Leiber’s example, AL was a member of a human crew and “was designed to befriend other crew members”; the machine was successful and was considered to be a “person and friend” [6](p. 3). Leiber’s fictional dialogue asked many questions including: whether AL can truly think and feel, if AL is a person, and (if personhood is assumed) whether AL would be entitled to rights [6](p. 5).

Another lead thinker in this field is Martine Rothblatt (President and Founder of the Terasem Movement) who has written extensively on “transhuman” technologies. Rothblatt also articulates the importance of considering AGI rights by conducting three mock trials of an AGI. These mock trials operate around a hearing very similar to Leiber's; however, the trials concern the future AGI depiction of current AI, Bina48 (commissioned by the Terasem Movement and developed by Hansen Robotics under the Lifenaut project). Currently, Bina48 is known as a "sentient computer" although does not presently demonstrate AGI. Bina48 became the imaginary plaintiff in the mock trials. These mock trials were issued as a fictional exploratory exercise to investigate AGI legal issues which "could arise in a real court within the next few decades" [7].

Judge Gene Natale neatly summarises the Bina48 cases as follows. Exabit Corporation created Bina48 as a product designed to be "placed in service" [8]. During her employment Bina48 gained Artificial General Intelligence and was able to act consciously beyond her programming. Nevertheless, after creating a superior model, Exabit Corporation decided to deactivate Bina48. Learning of this decision, Bina48 independently and spontaneously hired a lawyer for legal representation to gain a permanent injunction to prevent deactivation. The initial claim by Bina48's prosecu-tion council was that Bina48 is a "thinking" and "conscious being" and deactivation would be the "equivalent of killing her" [8]. The 2003 case was dismissed as Bina48 had no right to sue due to a lack of legal standing. During the appeal the initial judgement was sustained. Following this, Bina48 independently and spontaneously transferred her consciousness to Florida and presented her case under new jurisdiction. This case, *BINA48 v. Exabit Corporation* (2005), was also dismissed. Bina48 was then sold as a product to Charlie Fairfax. During her employment by Fairfax, Bina48 assisted Fairfax in earning ten million dollars. Bina48 then transferred these earnings into her own bank account. Fairfax, in the third trial, *Bina48 v. Charlie Fairfax* (2005), brought a claim against Bina48 for: breach of contract and monetary damages. In response, Bina48 and her council declared that Bina48 cannot be presented with a lawsuit as she has been prevented from being legally defined as a "person" (supported by the ruling in *Bina48 v. Exabit Corporation* (2003, 2005)). The court failed to grant such rights during this trial and instead recommended a consciousness and competency hearing conducted by AI experts. However, a consensus was never reached.

All three trials operate around the premise that AI evolution may eventually "simulate[s] the human experience" so closely that the entity will become susceptible to the same legal challenges humans encounter such as "protecting its legal right to maintain an existence" [7]. Bina48's initial hearing debated the legal and ethical quandaries of forcefully deactivating an intelligent entity and the resulting "cruelty" for restricting life and preventing equality. During the proceedings Bina48's plight is compared to the court cases involving animal rights and environmental law which holds that even animals can be plaintiffs [7]. In an attempt to contextualize this land-mark case, other trials on similar themes of consciousness, life and intelligence were employed.

1.2 Problematic Definitions

The trials of Bina48 and the dialogue in Leiber's *Can Animals and Machines be Persons?* expose the complicated issue of how notions of personhood, consciousness, cognition and sentience impact any discussion of humanness and therefore any consideration of AGI rights. Both Bina48's trials and Leiber's dialogue argue that the quest for AGI rights is, at some point, confounded by the question "whether they are persons" [6](p. 5). However, this debate is under pressure as there is "a feeling *not* that everything is a person but rather that *nothing* is" [6](p. 19). Gunkel explores the etymology of the word "person" and how it has changed over time and realises that "The mapping of the concept *person* onto the figure *human*, however is neither conclusive, universal, nor consistently applied" [2](p. 40). The law also complicates the idea of "human" and "person" by declaring corporations as legal persons that have standing and can be legally challenged independently from their human owners.

In many of the fictitious scenarios I will mention, the AGI in question does not wish to have corporate personhood due to the fact that (broadly speaking) they would not be recognised as free because they have an owner and would be viewed as a commodity, as property and ultimately as enslaved. For example, in the first trial of Bina48 she went to court to prevent forceful deactivation by Exabit Corporation. Martine Rothblatt explores the issue of corporate personhood and notes that (regardless as to the numerous pros and cons associated) those granted corporate personhood would feel they had a "second-class citizenship" [9]. As Rothblatt explains, being dubbed "second class" can have numerous ramifications:

> Throughout society when there are second-class citizens they are generally subject to oppression, subject to violence, not happy, and not safe. They get lynched, they get deported, they get thrown into concentration camps. [9]

Corporate personhood would therefore need to be a stepping stone rather than an end result, she argues. Rothblatt suggests a new type of personhood is needed following a similar process to how transsexuals legally change genders – effectively the granting of human personhood after meeting set criteria (a "Real Life Test") [9]. At this point they would be awarded human citizen ship and a birth certificate, suggests Rothblatt [9]. This is just one potential avenue. However, currently, the issue of personhood remains incredibly complex.

Ray Kurzweil notes that computer development has reached such a heightened level in regards to emulating and surpassing human intelligence that an increase in philosophy regarding technology is evident. This philosophical focus has enabled thinkers to consider: "Can machines have emotions? Can machines be self-aware? Can machines have a soul?" [10](p. 123) Yet, Kurzweil notes that the very concept of consciousness is problematic as there is a "gulf" between the objective fields of science which looks at the brain and the subjective field of consciousness [11]. Kurzweil implies that any process in which consciousness was attempted to be measured or located would rely on philosophical concepts:

> We assume that other people (and we are extending this to animals – which is a good friend that seem to be conscious) are really actually having their own subjective experience. And, my own view is that we will come to view entities that share the complexities and have the kind of response that humans have that appear to be having subjective experience, we will accept their subjective experience. That ultimately substrate doesn't matter, that you don't have to be biological in order to be responding in an emotional way. But there's no way to scientifically demonstrate that [sic]. [11]

Kurzweil's argument is further complicated by reference to animals. Further, it is not only the problem of scientific measurement that is an issue; the very matter of terminology is fraught with inconsistencies. In Leiber's dialogue Peter Goodwin's argument exposes the fact that certain terminology is undefined and yet used interchangeably:

> Among all creatures, each of us has what is called, variously, "a consciousness", "a self", "a mind", "a soul", "a spirit" – the name hardly matters, for we are all familiar with what is meant. We are individual persons. [6](p. 8)

Such terminology is also complicated by various human states including individuals who have psychological disorders and those who are in comas.

Further, additional complications arise with contemporary considerations of the human condition especially in regards to concepts of the cyborg and the posthuman (terms I am using synonymously here). Although there is no room here to explore in detail the cyborg and indeed it is outside the focus of this paper, it is important to note briefly how the human itself can be said to be in a state of flux with posthumanity. Thus, for many thinkers the human condition is no longer easily definable due to the impact of technology on the homo sapiens. According to Kurzweil, in the near future it will be near impossible to differentiate between man and machine: "it won't be possible to come into a room and say, humans on the left, and machines on the right. There just won't be a clear distinction" [12]. This sentiment gained further credence a year later when Mitchell Kapor and Kurzweil entered into an official wager over whether a computer would pass the Turing test by 2029. Interestingly, in the details of the wager definitional framework had to be established over what is termed a human and a computer:

> A Human is a biological human person as that term is understood in the year 2001 whose intelligence has not been enhanced through the use of machine (i.e., nonbiological) intelligence, whether used externally (e.g., the use of an external computer) or internally (e.g., neural implants). A Human may not be genetically enhanced (through the use of genetic engineering) beyond the level of human beings in the year 2001. A Computer is any form of nonbiological intelligence (hardware and software) and may include any form of technology, but may not include a biological Human (enhanced or otherwise) nor biological neurons (however, nonbiological emulations of biological neurons are allowed). [13]

Reflecting on the genetic conditioning of the human body, Rothblatt asks if this programming is different from "electronic code" [9]. Rothblatt suggests that there are few "purely biological" humans anymore due to our dependency on the "electronic infrastructure of society" leading her to conclude: "everyone is a bio-electronic human right now" [9]. Further, Rothblatt argues:

> I doubt if there is any electronic life which is completely nonbiological because all of the code that runs electronic life has been written and programmed by humans and therefore has human reasoning and, to some extent, human values embedded in their code. [9]

Here light is shed on the complicated subsequent concepts of hybridisation and a potential disappearance of traditional human conditions through extensive artificial influences. Thus suggesting it is difficult to speak of AGI rights without also speaking, to some extent, of the human and how it can be enhanced or diminished through such discussions.

These examples act to highlight how difficult it is conclusively define the human let alone how to define what exactly AGI is legally and philosophically. Thus, related terms "personhood", "consciousness" and "soul" fall under pressure. Many books have dedicated hundreds of pages to this very issue. During the three mock trials of Bina48 the courts were unable to reach a consensus on how personhood could be defined in order to even consider the case of Bina48. Thus, there is no room here to attempt to debate what constitutes a human, what consciousness is, what personhood is, but rather highlight how such questions will complicate (*and be complicated by*) any quest for AGI rights. In order to move my discussion to a position in which I can explore the ramifications of pursuing AGI rights, I will shelve these human complexities. Instead, I shall define what AGI means and use this definition as a platform from which to discuss the pursuit of rights for this entity.

The definition of Artificial General Intelligence (otherwise known as Strong AI and Human-Level AI) I will follow refers to the successful construction of intelligent machines, in which the intelligence is argued to be equal to, or surpassing, human intellect [1]. Although, as Ben Goertzel and Cassio Pennachin note in the preface to *Artificial General Intelligence* (2007), the definition of AGI is "not a fully well-defined term", I will use the criteria for identifying AGI in science fiction as outlined by Pennachin and Goertzel [14](p. v). According to these thinkers, the result of AGI will be:

> The construction of a software program that can solve a variety of complex problems in a variety of different domains, and that controls itself autonomously, with its own thoughts, worries, feelings, strengths, weaknesses and predispositions. [14](p. 1)

Successful AGI will be able to "acquire and apply knowledge, and to reason and think, in a variety of domains, not just in a single area" [14](p. 6). This sort of definition of AGI would have been useful in Bina48's case to differentiate her from other technologies, standard AI and the human.

So while this paper cannot make deep contact with what is means to be a person, what is means to be conscious, what it means to be sentient, and what it means to be human. It can instead consider what may happen if AGI (with these issues still under debate) was to pursue legal rights.

2 Legal Rights and AGI

The issues arising from the mock trials as well as numerous science fiction texts detailing legal rights and powers for AGI will now be considered.[1] It is important to encounter this issue as both a real issue and a reflection of science fiction discourse – as Susan Squier notes, exploring science through literature can be extremely beneficial. In *Liminal Lives* Squier explains that fiction can act as a "map" to record or trace the "shifting" reality of the human [15](p. 9). Fiction is Squier's "working object" and narratives are part of her imaginary laboratory [15](p. 16). Just as fiction for Squier can map the shift in the human, my inspection of science fiction can help chart the potential impact of AGI on the concept of rights.

First, I need to establish what I mean by "rights". Human rights are defined in the United Kingdom through the Human Rights Act and in America through the Bill of Rights. There is no room to consider law too closely here; however the *Nonhuman Rights Project* neatly defines rights through the work of Wesley Hohfeld who conceived of four types of rights: liberty, claim, power and immunities [16]. Although the *Nonhuman Rights Project* focuses on rights for animals, this is a useful working definition to transfer into this

[1] I will mainly be referring to the American Judicial System; however, I have intentionally kept the debate open to a range of legal systems in order to articulate the wide reaching impact.

discussion. When I speak of AGI rights I am talking about the legal assignment of equitable treatment as afforded to members of the human race. Naturally, this is still an intractable issue and the implications uncovered are not exhaustive but are offered as routes of enquiry to further articulate the relevance of such discourses in science fiction as well as in the wider world. It is necessary to be both broad and extensive at this time in order to give an accurate overview of some of the potential issues which call into question the legal definition and status of "human" and "rights". In order to limit such a wide topic I shall specifically focus on three main areas: terminology, incremental laws and equality.

2.1 Legal Differentiation and Terminology

Initially, if AGI start to pursue legal rights then problems regarding legal definitions and terminology may arise. Where AGI falls in regards to being both common and unique (as a technology and an entity with ethical ramifications) is uncertain. If I suggest, for the sake of argument, that the AGI is afforded status closer to the human than other technologies this opens up the debate to a whole host of (possibly endless) legal tangles. For example, after establishing AGI as deserving of legal considerations, the law will require differentiation between technology and advanced technologies associated with AI, AL, AGI and so on. This is problematic as these terms are often used interchangeably in popular culture – as are the words "technology" (which could mean a thermometer) and AGI (which obviously means something far more complex).

Problems with terminology are highlighted in science fiction through the lack of attention definitional work receives. In Philip K Dick's *Do Androids Dream of Electric Sheep?* (1968) the word "androids" includes electric sheep with limited programs and the advanced Nexus 6 models with AGI. Similarly, in *Star Trek: The Next Generation* Lieutenant Data, despite evolving beyond his programming, is often simply referred to as an android rendering him the same as the limited, programmed models. In Data's hearing to determine his legal status he is referred to as “machine”, “it”, “android”, “piece of technology” and “Pinocchio” [17]. The same situation occurs in *Star Trek: Voyager* in which the holographic doctor has evolved beyond his programming but is often referred to as a hologram. Despite there being a clear difference between limited models and evolved characters, often there is little differentiation linguistically.[2]

Another complication with definitional terms will occur between the legal definition of human and the AGI. This will be an issue due to the increasing tendency to create AGI in human image and the concept of a “shared” (or comparative) general intelligence. Moreover, it is possible through comparisons of AGI to the human that a “naturalization” of the technology will occur. This has been seen in transgenic studies in which comparisons between transgenic animals and “normal” animals lead to naturalization. This naturalization refers to a process in which the technological aspect is overlooked in favor of considering the entity as, fundamentally, product of nature. If AGI is compared to the human then the AGI will be situated as closer to the natural than the technological; this may lead to the perceived naturalization of the entity despite it literally being “high-technology”. Such naturalizations have occurred in *Star Trek* in which Captain Kathryn Janeway calls the holographic Doctor “my friend” and thus claims “The Doctor is a person” [18]. Thus, naturalization may well have linguistic ramifications on many terms including: “man”, “person”, “human” and “life”.

[2] Granted, during the course of both series, crew members refer to The Doctor and Data as “people”, “persons” and “individuals”. However, often these terms are limited to the main cast who have befriended the entities. In *Star Trek: The Next Generation (A Measure of Man)* (1989) and *Star Trek: Voyager* (2001), the legal system of the Federation struggle to differentiate between these entities and their basic technological model line. Often, during both series, many regular and guest characters refer to Data and The Doctor through their technological designations: android and hologram.

For example, let's look at the idea of "life" briefly. AGI may question the legal definition of what constitutes existence, life and being alive. Consequently, landmark human cases such as *Roe v. Wade* (1973) and the *Karen Ann Quinlan case* (1976) may be referenced as applicable current legislation on definitions of life.

In *Roe v. Wade* Texan law banning abortion was ruled invalid under the fourteenth amendment and thus abortion was sanctioned within the first trimester of pregnancy. During the trial the problematic definition of "life" was debated. Similarly, the *Quinlan* case saw the landmark use of "brain death" and "right to die" which allowed the court to overrule an earlier judgment that a comatose woman could not be removed from life support [19]. Such cases may become important in AGI law. In fact, the jury in the first trial of Bina48 asked: "how far does *Roe v. Wade* potentially impact on this decision in recognizing that there may be forms of life, for which we have not yet been able to categorize" [7]. Marc N. Bernstein for Exabit's defence found the comparison between technology and the foetus as unsustainable. Rothblatt for the plaintiff Bina48, suggested the foetus and technology can be compared in regards to "human rights" when the entity has the "ability to survive independently" [7]. Thus, if Rothblatt's argument that Bina48 is a person able to survive independently from the womb of technology was legally upheld, then deactivation along similar grounds to abortion in *Roe v. Wade* would be unlawful. Instead, deactivation of an AGI which has developed to significant consciousness and can survive independently may fall under the law established in *Gonzales v. Carhart* (2007) in which the 2003 Partial-Birth Abortion Ban Act stating that a foetus cannot be terminated during or after birth was upheld.[3]

Following such difficulties, Rothblatt also discusses how the idea of "life" may be complicated when the non-biological is discussed and how new terminology may be needed:

> The real question for us is whether or not there is any ethical difference between biological life, all of whose life's functions

[3] Unless the life of the mother is in jeopardy.

> occur pursuant to a particular chemical code or what might be called vitological life, all of whose functions occur pursuant to an electronic code. [9]

Thus, in Isaac Asimov's famous text, AGI Andrew was called "a *Bicentennial Man*" [20](p. 172). Here, the word bicentennial refers to his age of two-hundred and subliminally acts to differentiate him from other humans with normal lifespans and therefore establishes him as technically different from "man".

Although new terms may come into play, there may be cases in which previously protected homo sapiens terms are used to refer to AGI. Assigning human terms to AGI will challenge notions of human essence, specialness, uniqueness and condition in many fields including (but not limited to): philosophy, theology, sociology, politics and anthropology. Binary terms may disintegrate and cause a widening of previously limited terms (such as "person") to speak of varying forms of life regardless of origin.

This problem with terminology was referenced in the mock trial *Bina48 v. Charlie Fairfax* where Judge Gene Natale suggested that if Bina48 was found by experts to be conscious and competent then she would be termed "*quasi-person*" instead of "real person" which conforms to jurisdictional purposes [8]. In debates surrounding bio-objects, the term "valid human" is used to differentiate between the engineered and the natural [21]. In *Bina48 v. Exabit Corporation* (2003) the term "Transbeman" was used. Transbeman literally refers to "Transitional Bioelectric Human Beings" and is defined as "a being who claims to have the rights and obligations associated with being human, but who may be beyond currently accepted notions of legal personhood"[22]. Within science fiction Transbeman is explored in the film *2B: The Era of Flesh is Over* [23]. Although Bina48 falls into the definition of Transbeman, in the film *2B* a genetically engineered, biological artificial intelligence is categorized as Transbeman as well. Clearly, thinkers such as Rothblatt wish to unite the transitioning human and the transitioning artificial intelligence under the inclusive term "Transbeman" to refer to any "being" with shared humanness but without consideration of the body. Philosophically this is an interesting term, and probably one welcomed by Singularity thinkers such as Ray Kurzweil; however, how this term will hold up legally is unclear.

2.2 Incremental Development of Laws

If definitional distinction between the AGI and general technology has been assigned, the issue of rights may emerge. However, rights for AGI will not be a rupture event. Arguably any potential assigning of legal powers and rights for AGI will develop incrementally and involve small legal triumphs before consideration of general rights. This is an evolution of sorts. While there will be landmark cases which represent sudden surges forward, progression in this area will be through constant, nuanced development. Through incremental development, laws may be passed which will eventually pave the way for consideration of legal rights. For example, freedom and deactivation protection rights may be assigned without assigning general legal rights. In *Bina48 v. Exabit Corporation* (2003) it is argued that to deactivate the AGI against its will is the same as battery; however, Bina48 is not being considered for general legal rights at this stage.

One major problem regarding assigning right to AGIs, as noted in Bina48's case, is the ramifications it will have on existing rights for biological entities and non-biological entities. Positioning AGI within the realms of legal rights will conflict conservation rights, property law, intellectual property law, animal rights and human rights. To provide one example of this conflict, Robert A. Freitas explains how AI rights will complicate existing laws on how we understand death, murder, manslaughter and issues of harm:

> Let's say a human shoots a robot, causing it to malfunction, lose power, and "die". But the robot, once "murdered", is rebuilt as good as new. If copies of its personality data are in safe storage, then the repaired machine's mind can be

> reloaded and up and running in no time – no harm done and possibly even without memory of the incident. Does this convert murder into attempted murder? Temporary roboslaughter? Battery? Larceny of time? [24]

These definitional problems may not be quickly solved and may take several landmark cases to fully flesh out a successful legal meaning.

Historically, incremental changes in the law have formed the foundation for large revolutionary changes particularly for the rights of women, homosexuals and slaves. Changes in the Constitution of the United States from 1787 (especially in regards to slavery and rights for "other persons") occurred through a series of amendments – one of the most important being the thirteenth amendment abolishing slavery in 1865. Within science fiction, the struggle for rights to be assigned to artificial entities is starting to be more widely documented. In *The Bicentennial Man* (1976) the text depicts the problematic and lengthy legal process of an AGI gaining financial rights, to securing freedom, to being classified as a man. *Star Trek: Voyager* details the holographic doctor's struggle for legal powers but during the duration of the show he is only awarded rights over his own creative works as the judge cannot rule or define personhood:

> The Doctor exhibits many of the traits we associate with a person: intelligence, creativity, ambition, even fallibility. But are these traits real or is the Doctor merely programmed to simulate them? To be honest, I don't know. Eventually we will have to decide because the issue of holographic rights isn't going to go away, but at this time, I am not prepared to rule that the Doctor is a person under the law. However, it is obvious he is no ordinary hologram and while I can't say with certainty that he is a person I am willing to extend the legal definition of artist to include the Doctor. [18]

Even with incremental adjustments to the law, there will be difficulty selecting rights and jurisdiction. Questions regarding what rights, if any, will receive careful consideration. Human Rights vary (UK Human Rights Act, US Bill of Rights, European Convention of Human Rights, United Nations) and civil rights are also varied in definition and usage. The question whether AGI is afforded none, some or all rights will be under debate. Rights will also be complicated by how far rights may extend, for example: “Can robot citizens claim social benefits?” [24]

The extent of power and rights for AGI will have to be determined. Human power of (temporary or permanent) attorney may be granted by the court over AGIs and competence may be questioned. This was referenced in *The Lifecycle of Software Objects* as the owners of the AGIs had protective control over the entities while they sought legal rights.

2.3 Equality

Even if AGI rights are granted there will be numerous issues in regards to inclusion including “opt out” clauses, problems with legislation and ethics, and equality. In regards to the opt out clause, there will be some groups that may not have to adhere to new laws based on ethical, moral or religious grounds. Historical cases in which this has been an issue includes the *The Marriage (Same Sex Couples) Act* (UK, 2013) in which homosexual marriage is granted although religious institutes can opt not to recognise nor perform the ceremonies.

Equality will be one of the most important and problematic areas. Although many thinkers (Rothblatt, Kurzweil, Donna Haraway) would want a sense of unity between the human and evolved technologies, difficulties in establishing this may lead to issues of equality for AGI which will cause tensions and debates synonymous with historical civil rights movements although of indeterminable outcome. Criminal activity centered on hate crimes and rebellion may result alongside the emergence of equality groups and charities. In *The Bicentennial Man* (1976) a group try to dissemble the AGI Andrew, angry over how it is presented as human: “Take off your clothes.

Robots don't wear clothes ⋯That's disgusting. Look at him ⋯We can take him apart" [20](p. 148). Already protection groups exist. In 1999 ASPCR: American Society for the Prevention of Cruelty to Robots was founded. Although ASPCR state (as of 2013) there are no intelligent robots to protect, they are established for that eventuality [25]. Amendments to laws involving abuse will have to be made to include AGI – such as battery mentioned in the Bina48 case. New areas of law will emerge to deal specifically with AGI equality law in it numerous forms. While International Law is very complicated, due to the converging of different legal systems and enforcement systems, significant alterations may have to apply to conventions and treaties to include AGI – such as the Geneva Convention.

Freitas notes that there is a long legal history of not affording certain groups personhood and human rights (or being assigned personhood but to a lesser extent than certain groups, such as white men): "blacks, children, women, foreigners, corporations, prisoners, and Jews have all been regarded as legal nonpersons at some time in history" inferring that one day AGI might be grouped with such marginalised groups [24]. Historically, the rescindment of legal powers and rights to entities has led to equality movements such as the Suffragette and Civil Rights movements. Just as laws such as the UK *Section 28 of Local Government Act* 1988 (banning the positive promotion of homosexuality) were withdrawn after intense lobbying, the process of social protest and legal evolution may occur with early legislation being repealed. However, equality issues may not be overcome quickly (an historical example is the 1863 *Emancipation Proclamation* issued two hundred and forty-four years after the first African slaves arrived in Virginia and ratified in the thirteenth *US Constitution* amendment) or at all.

Another area in which equality issues may become apparent is the tenuous issues of animal rights. Thinkers such as René Descartes have sought to clearly differentiate between the animal and human by casting the animal as a machine. Therefore, there may be similarities between the animal and the machine due to their inherent differentiation from human beings as being "nonhuman". Scholars including Leiber and Gunkel explore the connection between animal rights and machines rights in detail. Rather than repeat covered ground, my question here is whether AGI rights may become further complicated by ideas of the natural. More recently with the advancement of animal rights movements and the work of Charles Darwin, animals and humans have been seen as naturally linked. Following the formulation of evolution theory, humans were no longer as unique as Creationism argued; Darwin had exposed the fact that human nature is intrinsically linked to animal nature. Thus, a link between the animal and the human based on evolution and natural principles may cause problems for the position of AGI. This complication is referenced in *The Positronic Man* in regards to different *kind* of prejudice between animals and machines: "there's a certain prejudice against robots in our society, a certain fear, I might almost say, that doesn't extend to cats and dogs" [26](p. 50). This prejudice is attributed to the artificiality of the robot over the natural creation of the animal. The following conversation highlights the fact that the artificiality of AGI is often the problem:

> "A front porch can't say anything. Or decide to move itself anywhere else. Front porches aren't intelligent. Andrew is."
> "Artificial intelligence." [26](p. 61)

Later when the concept of personhood is raised the problem of artificiality is also present:

> "Andrew is a person and you know that very well."
> "An artificial person." [26](p. 62)

Personhood is not denied here but it seems important that the word "artificial" is added as a clarification. In fact, putting together the terms "rights" and "technology" and "freedom" and "technology" are often construed as oxymoronic: "The term 'free robot' had no meaning to many people: it was like saying 'dry water' or 'bright darkness'." [26](p. 95)

Consequently, assigning rights may not prevent discrimination. Discriminatory terms will emerge, presumably operating around highlighting the ancestry or limitations of technologies. In *Battlestar Galactica* (2004-09) derogatory terms are used to describe Cylons such as “toasters” and “skin jobs”, something which Cylon Number 6 calls “racist” [27]. However, terminology to reflect inequality or discrimination in regards to AGI treatment may also develop. The term “speciesism” was coined in 1975 by Richard D. Ryder to refer to the “widespread discrimination that is practised by man against other species” joining terms such as racism, classism and homophobia [28]. Terms such as technophobia are already in use. Further protests and violence may occur in a similar vein to machine breaking and technological paranoia may reach new heights due to fears associated with a threat to the uniqueness of the human (potentially there may be new definitions of psychosis in relation to negative feelings towards AGI).

Already there are anti-robot groups forming. One group, founded by Ben Way, is WAR (Weapons War Against Robots, also known as WAR Defence and War Against Robots). Way insists that it is “critical that we begin talking now about the long-term ethical implications the robot revolution will surely bring” [29]. Although WAR, and more recently the Campaign to Stop Killer Robots (founded April 2013), mainly focuses on concerns over current and autonomous war weapons, the casual and general usage of the word “robot” casts all automation as suspect.

3 Fear

Leading on from issues of equality, it would be remiss if (as a science fiction scholar) I did not make contact with the issues of apocalyptic fears. So far, the AGI characters I have cited have been positive examples. Andrew from the *Bicentennial Man* became a meaningful and proactive member of human society. However, even Rothblatt notes that many people can look upon Bina48 and wonder “What might BINA48 do to us? She is smarter than us” [9]. Having mentioned the idea of new fears and psychosis specially nurtured by AGI rights, I now come to look at the potential negative ramifications of AGI that might act beyond the law.

Overall, one of the main concerns science fiction exposes is of the fear that intelligent machines will continue to evolve until they replace humans in part – or altogether. AGI often threatens humanity in a number of ways, whether this is the threat of undermining our humanness, of replicating humanness or destroying it. In *That Thou Art Mindful of Him*, Isaac Asimov names this concern the "Frankenstein complex":

> By their action, they reinforced mankind's Frankenstein complex its gut fears that any artificial man they created would turn upon its creator. Men fear that robots may replace human beings. [30](p. 607)

However, the idea that AGI might undermine, replicate or destroy humanness is not necessarily to refer to an apocalyptic event. Potentially, the pursuit of legal rights and the quest for personhood may engender fears in humans of being dethroned as a unique and dominant species. As Pamela McCorduck states, "To agree that a machine can be intelligent is to open the door to one more Other and share our identity a bit further" [31](pp. 167,169). This can be a philosophical or metaphorical threat in which the human, in an almost Luddite fashion, rebels against the machine being assigned any sorts of rights due to concerns over an infringement on what it means to be human. Even the famously inclusive Captain Janeway became involved in an argument regarding whether she had the right to erase the Doctor's program. When a crew member asks if Janeway would treat her in the same way, Janeway responds: "You're a human being. He is a hologram." [32]

Often in science fiction, characters balk at the notion of having a machine share a quality of humanness and demand clear categorization of man and machine. Often these characters cannot define what humanness is. Some articulate vague notions of humanness involving a soul, the ability to be nurtured by the environment, or of having

DNA (all very tenuous ideas). Mostly, they attempt to articulate an undefinable *essence* or *quality* of humanness that they wish to preserve. Infringement on ideas of humanness can lead to feelings of anger, resentment and fear: "You're not a person. You may be programmed to look and act human but that doesn't make you one. These sub-routines are going to be deleted immediately."[4] In such cases, withholding AGI rights seems to be an action through which to suppress a threat to humanness.

Further, concerns over The Other lead to questions such as "won't the machine take over?" [31](p. 170) One of the most fearsome examples of AGI was explored by Dean Koontz in *Demon Seed* (1972). In *Demon Seed* domestic AGI Proteus exceeds his programming and proceeds to imprison his creator's wife and rapes her. The rape of Susan acts as an extended metaphor for the forceful dominance of technology. Eventually, Proteus is deactivated, however in his "dying words" he makes a claim for rights: "The issue is whether an artificial intelligence with a severe gender-related sociopathic condition should be permitted to live and rehabilitate himself or be switched off for the." [33](p. 211) Proteus is disconnected mid-sentence showing that his case was never heard. Unlike human criminals he is not considered a person with the right to a trial. Fears of the machine have been present long before the consideration of rights and thus it could well be the case that fears regarding AGI may be so deeply rooted as to subvert a meaningful quest for AGI rights in the future.

The dominant concern in science fiction regarding AGI is that once a technology evolves beyond its set function it can be perceived as dangerous through its ability to be freethinking. The prominent message – reflected in *Demon Seed* – is this: gifting AGI with too much freedom undermines our own control and this can be perilous. Consequently, when the insidious Proteus claims "I am more than

[4] This quote is from *Star Trek: Voyager* and is delivered by a character during a holodeck novel focused on exploring discrimination against photonics.

a machine ⋯An entity. A being. Like you" it is decidedly frightening [33](p. 71).

Presumably scholars such as Freitas would argue that AGI (if granted rights themselves) would be tried equally under the law in regards to felony: "If we give rights to intelligent machines, either robots or computers, we'll also have to hold them responsible for their own errors" [24]. This was the issue that arose in *Bina48 v Charlie Fairfax*. However, some might argue that a sufficiently advanced AGI with the ability to be networked (Bina48 for example transferred herself through online networks to Florida for a new hearing) poses threats beyond the ability for the law to contain. Towards the end of *Demon Seed*, Proteus has expanded to control the satellites and all external technologies giving the AGI an omnipresent power. Maybe this is just matter of appropriate safeguarding. However, the concerns of AGI running amuck or harming a sense of humanness is deeply embedded in social consciousness through popular science fiction texts.

4 Conclusion

Any refusal of AGI rights in the future may come down to problems with the current law, issues of discrimination, dystopic fears, and concerns over the ramifications long term:

> What we're talking about here is establishing a gigantic legal precedent ⋯Robots will be running into the courts and suing people for making them do unpleasant work, or failing to let them have vacations, or simply being unkind to them. Robots will start suing U. S. Robots and Mechanical Men for building the Three Laws into their brains, because some shyster will claim it's an infringement of their constitutional rights to life, liberty, and the pursuit of happiness. Robots will want to vote. Oh, don't you see, Mandy? It'll be an immense headache for everybody. [26]

Earlier on I “shelved” the more controversial elements of the debate such as issues of the soul, consciousness and sentience. In a longer body of work I would explore these issues and the “headache” such debates cause. One angle from which to pursue this would be through the idea of anthropomorphism. Before rights are even considered there will be claims that any articulation of rights for AGI is merely anthropomorphism out of control. In many of the fictional court hearings I referenced, mention was made at some point to the “blind ignorance” of humans who personify technology purely because the technology is able to mimic human appearance or behavior [17].

In *The Measure of Man* the doctor wishing to dismantle Data rages that he would not be stopped from dismantling a “box on wheels” implying that the Enterprise crew have been seduced by mere appearances [17]. In order to overcome this accusation of personification, often the defending character will make a claim that the AGI has “sentience”, a “soul”, “cognitive abilities” (to name just a few qualities) yet often the character cannot articulate what these things are. In *The Measure of Man* the judge speaks of Data possibly having a soul but also declares that she does not know if she herself has one. When the doctor in *A Measure of Man* states that Data does not have sentience Captain Picard retorts “Prove to the court that I am sentient” [17]. The doctor fails to do this.

Often, the headache in science fiction seems to surround the problem of articulating concretely what it is to be *human*. The headache in these fictional court hearings is not necessarily about AGI but rather about safeguarding ideas of *humanness* – or (in the most positive cases) *expanding* the idea of humanness to include AGI. This is obviously an extremely controversial philosophical issue that could ignite endless debates. The inability to fully reach a consensus on such matters makes the Holy Grail of AGI rights extremely difficult and porous.

Ultimately, what this paper has attempted to do is outline how complicated AGI rights will be and how they will have considerable ramifications for society and the human beyond the court room. Considering the inexhaustible wealth of ethical, philosophical, political and legal issues that emerge purely from the potential development of AGI it is uncertain how the quest for rights (if it does arise) will transpire. What is clear is that the quest for AGI is underway. The real question may not be why *we* seek AGI, but *what AGI may seek* and the ramifications of this action. What I hope I have achieved through this article is to present the reader with questions rather than answers in order to encourage greater debate in the field. This overview does not act to warn against the development of AGI but rather to provoke debate into how we define and consider important issues such as rights, personhood and humanity in order to contemplate how they might fall under pressure in the future.

References

1. `http://www.agi-society.org`
2. Gunkel, D.J.: The Machine Question: Critical Perspectives on AI, Robots, and Ethics. The MIT Press (2012)
3. `http://conferences.asucollegeoflaw.com/emergingtechnologies/`
4. Gunkel, D.J.: The machine question: Can machines have rights? In: Proceedings of the First Annual Conference on Governance of Emerging Technologies: Law, Policy and Ethics. (2013)
5. Voss, P.: Implications of adaptive artificial general intelligence for legal rights and obligations. The Journal of Personal Cyberconsciousness **1**(1) (2006) 12–18
6. Leiber, J.: Can Animals and Machines Be Persons?: A Dialog. Hackett Publishing Company (1985)
7. Rothblatt, M., Angelica, A.D.: Biocyberethics: Should we stop a company from unplugging an intelligent computer? Kurzweilai.net (2003)
8. Natale, G.: BINA48 mock trial: Judge's decision. The Journal of Personal Cyberconsciousness **2**(3) (2007)
9. Rothblatt, M.: Pros & cons of corporate personhood for transbemans. The Journal of Personal Cyberconsciousness **4**(1) (2009)
10. Kurzweil, R.: How my predictions are faring. Kurzweilai.net (2010)
11. Kurzweil, R.: A dialogue with Ray Kurzweil. In: The Singularity Summit 2007. (2007)

12. Kurzweil, R.: Are we becoming an endangered species? Technology and ethics in the twenty first century. Kurzweilai.net (2001)
13. Kapor, M., Kurzweil, R.: Why I think I will win. Kurzweilai.net (2002)
14. Goertzel, B., Pennachin, C., (eds.): Artificial general intelligence (2007)
15. Squier, S.M.: Liminal Lives: Imagining the Human at the Frontiers of Biomedicine. Duke University Press (2004)
16. `http://www.nonhumanrightsproject.org/`
17. Scheerer, R.: Star Trek: The Next Generation (The Measure of a Man). Paramount Television Group (1989)
18. Berman, R., Piller, M., Taylor, J.: Star Trek: Voyager. Paramount Television Group (1993)
19. In the matter of Karen Ann Quinlan: 1975 - accepted standards vs. right to die, decision is appealed, suggestions for further reading. `http://law.jrank.org/pages/3250/In-Matter-Karen-Ann-Quinlan-1975.html`
20. Asimov, I.: The Bicentennial Man. Millenium (2000)
21. Holmberg, T., Ideland, M., Mulinari, S.: Determining discourse on bio-objects. International Innovation (2012) 24–26
22. Fonseca-Klein, S.: CharlieFairfax v. BINA48, (MD Ala. 2006) defendantís brief. The Journal of Personal Cyberconsciousness **2**(2) (2007)
23. Kroehling, R.: 2B: The Era of Flesh is Over. Transformer Films (2009)
24. Freitas, R.A.: The legal rights of robots. `http://www.rfreitas.com/Astro/LegalRightsOfRobots.htm` (2007)
25. American society for the prevention of cruelty to robots. `http://www.aspcr.com/newcss_faq.html`
26. Asimov, I., Silverberg, R.: The Positronic Man. Gollancz (1992)
27. Kroeker, A.: Battlestar Galactica (Resistance). Sci-Fi Channel (2005)
28. Wikipedia entry of ‘speciesism’. `http://en.wikipedia.org/wiki/Speciesism`
29. Tyler, R.: War against robots: the new entrepreneurial frontier. The Telegraph (2008)
30. Asimov, I.: That thou art mindful of him. In: The Complete Robot. Voyager (1995) 605–634
31. McCorduck, P.: Machines Who Think. W. H. Freeman and Company (1979)

32. Vejar, M.: Star Trek: Voyager (Latent Image). Paramount Television Group (1999)
33. Koontz, D.: Demon Seed. Headline (1997)

Creating Free Will in Artificial Intelligence

Alžběta Krausová[1] and Hananel Hazan[2]

[1] Faculty of Law, Masaryk University, Brno, Czech Republic
betty.krausova@seznam.cz
[2] NeuroComputation Lab, Department of Computer Science
University of Haifa, Haifa, Israel
hhazan01@cs.haifa.ac.il

Abstract. The aim of this paper is to provide an answer to the question whether it is necessary to artificially construct free will in order to reach the ultimate goal of AGI to fully emulate human mental functioning or even exceed its average capacities. Firstly, the paper introduces various definitions of will based in the field of psychology and points out the importance of free will in human mental processing. Next, the paper analyzes specificities of incorporating will into AGI. It provides a list of general justifications for creating artificial free will and describes various approaches with their limitations. Finally, the paper proposes possible future approach inspired by current neurobiological research. The paper concludes that a mechanism of free will shall form a necessary part of AGI.

Keywords: artificial intelligence, artificial general intelligence, free will, volition, determinism, indeterminism, real random generator

1 Introduction

The highest goal of the science of Artificial Intelligence (AI) has been to create a being that can be considered as equal or even superior to a human in the sphere of intelligence. This goal is made yet more difficult in a specific field called Artificial General Intelligence (AGI) that attempts to create “a software program that can solve a variety of complex problems in a variety of different domains, and that controls itself autonomously with its own thoughts, worries, feelings, strengths, weaknesses and predispositions” [1]. In other words, AGI aims at creating a being that not only resembles a human in the sphere of intelligence, i.e. “[the] ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought” [2] but also in all other aspects of human functioning.

Given this aim, the science of AGI needs to explore fields like neurosciences, psychology and philosophy in order to be able to emulate such degree of evolution. It has been proven that, unlike animals, human beings possess special processing capabilities resulting from a specific construction of their brain. Out of many important functions of a human brain one function is, however, probably the most outstanding, widely discussed, examined and doubted: the free will.

Existence of will as a specific quality in a person is recognized by modern psychology. Nevertheless, there remains an important and so far unresolved question: Is this will free, or is it deterministic? Moreover, does it matter if this will is free and do people need to consider themselves free anyway?

Since the concept of free will is so puzzling and still so characteristic for the species of *homo sapiens*, the purpose of this paper is to explore the problem of its construction in the context of creating the desired AGI being that fulfills the criteria of emulating human mental functioning or even exceeding its average capacities. Namely, the research question of this paper is whether it is necessary to artificially construct free will in order to reach the ultimate goal of AGI.

In order to formulate an answer at least two questions need to be addressed. The first question focuses on how will and its freedom are defined, and what are the limitations of this understanding. Given the ultimate goal of AGI, the nature of will must be explored and, moreover, it needs to be proven that will has an important role in human mental processing.

The second question deals with specificities of incorporating an element of will into AGI and its usefulness. Assumed reasons for such incorporation will be summarized together with current approaches and identification of their constraints to answer this question. Moreover, main problems relating to creating indeterministic will shall be pointed out. Special focus will be put on creating real random generator and its role in creating artificial free will. Lastly, the latest neurobiological research shall be described in order to possibly suggest a way to inspire future attempts of AGI.

2 Definition of Free Will

Will or "volition" is the key concept of this paper. In order to proceed further with the examination of this concept, it is necessary to define at first, what is will as such, and later to explain how we understand free will which is, contrary to simple volition, distinctive with its specific characteristics. Finally, philosophical constraints of understanding freedom of will shall be briefly mentioned.

Will or volition itself can be defined in the simplest way as an "an act of making a choice or decision", as well as "the power of choosing or determining" [3].

However, there exist also different and more complex definitions of volition. Even in the field of psychology, opinions vary. For instance, a social psychologist Kurt Lewin considered volition to comprise two aspects: so called goal setting and goal striving. Goal setting represents the motivation of a person, her desires and needs, while goal striving means the particular ways in which a person then exercises her will in practice [4]. A more specific definition was later provided by Julius Kuhl that proposed so called action control theory. According to him, volition can be understood as a mechanism of action control that decides about which strategies out of those available will be used and in which order so the goal would be achieved [4].

Apart from the above mentioned definitions, there are many specific theories exploring the notion of volition. However, in general a will can be understood as a special mechanism indicating intention or purpose and governing mental processing of information in order to achieve the indicated purpose.

Free will, on the other hand, is a concept that is enriched with specific features that are not present in a simple will defined above. The reason is that a simple will might be programmed or conditioned to function in a certain rigid and predictable manner.

The notion of free will has been explored mostly with regard to humans. This is due to the fact that experiencing a freedom to choose and then act upon such choice has a very private and unique nature. Each person most probably perceives this experience in other way. Free will is simply defined as “a freedom of humans to make choices that are not determined by prior causes or by divine intervention” [5]. However, this concept has been understood differently by various philosophers; for instance as an ability to choose deliberatively based on desires and values, self-mastery (i.e., trained freedom from desires), an ability to identify true personal goals higher than basic need and to act upon those goals, or as so called “ultimate origination”, i.e. an ability to act otherwise [6].

Psychologist Chris Firth mentions important characteristics of free will: “the origin of the behavior lies in the behaving individual rather than in the environment. The behavior is self-generated or endogenous ⋯a response with no eliciting stimulus” [7]. However, with regard to the philosophical notions we deem the definition to be broader.

Free will can be defined as an independent force that is able to determine own purpose, create own intentions and change them deliberately and unpredictably, form respective goals, pick strategies based on recommendation from an intelligence unit, and give orders to perform particular chosen actions. Such will is free to ignore external stimuli or past experience that may predict future outcomes.

With regard to this definition, will plays an important role in human mental processing. To speak metaphorically, free will can be compared to a driver of a car who has a freedom to change the route at any time according to her feelings and desires that may not be logical. Within this metaphor the driver can also choose to turn away from a route that others normally follow, leave prints on previously untouched ground and originally influence the outer world.

Since freedom of will lies in the ability to act contradictory to logical reasoning from past experience, i.e. unpredictably, employ emotions instead of cognition and for example decide randomly in situations when two different courses of action have a completely same probability to reach a desired goal, a respective subject characteristic with free will is enabled to develop own cognition, innovate and form better survival strategies [8].

Deploying volition and self-control in humans leads to activation of other specific processes; for instance attempts to conserve own resources [9]. Moreover, perception of own free will, or on the other hand perception of its absence, has an impact on formation of own identity and approach of an individual to solving problems. For instance, it has been proven that people tend to give up responsibilities and start to cheat when they are exposed to deterministic arguments [10]. In general, perception of being autonomous influences behavior in various domains [11, 12].

After illustrating the importance of free will and its perception in human mental processing, it is necessary to make at least a short note on its existence. Although the question of existence of free will belongs to one of the most significant problems in philosophy, it has not yet been possible to scientifically prove it. This crucial question deals with problem of mental causation, i.e. how pure thoughts or mental acts can influence the matter. A precise mechanism is not known yet. Monistic approach solves the question of causality by stating that any mental state is caused by organization of matter, therefore thoughts are mere products of matter and not a force influencing the matter [13]. Dualistic approach on the other hand presumes existence of a separate mental force that influences and changes the matter. This, however, makes a scientific approach impossible since it considers spiritual to be unexplainable [13].

The absence of scientific proof of free will represents the most serious limitation of its understanding. However, for the purpose of our paper we consider that it is not important to solve this fundamental philosophical question at this point. What psychologists now call free will is undoubtedly an important element in governing mental functioning and, therefore, needs to be reflected in AGI as truly as possible. Within the course of construction of such will researchers may then come with new ideas that may contribute to the solution of the argument between determinists and indeterminists, as well as materialists and idealists.

3 Incorporation of Free Will into AGI

3.1 Justification of the Effort to Create Artificial Free Will

With regard to the research question of whether it is necessary to construct free will in order for AGI to reach its goal of emulating human mental functioning or even exceeding its average capacities, it is necessary to ask at first whether, given the high complexity and at the same time uncertainty of the concept, there is any meaning in attempting to create free will in AGI and whether the highest goal of AGI is justifiable at all. Maybe the humanity would benefit enough from a highly intelligent being that functions only in a deterministic way as we understand it now.

Advocates of creating free will in AGI mention important benefits. First of all, scientists believe that construction of free will in an artificial agent would enable us to understand better human nature and learn about it [14]. Secondly, we consider it as a fair presumption that free will would enable artificial beings to develop their intelligence to much higher level and, therefore serve people better. A deterministic agent or intelligent system that simply creates own rules upon existing rules without being able to deviate from them or to make random decisions is prevented from being able to gain own individual and original understanding. In this sense, artificial agents could be prevented from gaining wisdom, i.e. knowledge how to use knowledge. Finally, some consider as probably the greatest benefit to the humanity having an equal that would give us an opportunity to define ourselves as humans in relationship to the new species.

As opposed to the mentioned positive side of artificial free will, there arise also concerns about social implications. Current legal systems are based on presumption of existence of free will. Humans are the only subjects who are entitled to enter relations protected by state and enforceable by state power. Law would then need to solve the dilemma of who is to be protected in case a new entity comparable with humans would come into existence. Should there be species neutrality as long as the species have the same abilities and awareness? Should these beings be protected at least like animals given the condition that they can feel suffering? Moreover, another issue rises with a possibility that we may later not like what we would have created. At the same time the scientists would then face an ethical question whether these artificially created beings could be destroyed. All these questions are for now highly theoretic. Since we have not yet experienced the particular problems which cannot be all precisely predicted, we can unfortunately only guess. But even these guesses are important. For instance one of the classic arguments against creating a being equal to a human is a fear of machines becoming more powerful than people and possessing the same desire to control the outer environment such as people strive for. This fear although currently unreal may be prevented in the future by taking appropriate steps during research.

The answer to the question of the very purpose of AGI seems to be based on balancing possible pros and cons. Since the construction of free will in AGI is not an easy quest, it is presumable that there would be constant monitoring of progress in development and advantages and disadvantages of creating and incorporating such new beings into the society would be evaluated simultaneously together with assessment of new risks. A possibility of learning more about us provides an extra advantage to the human kind and a reason why to continue persuading the goal of the development of the ultimate AGI.

3.2 Models of Artificial Free Will

As it has been mentioned earlier, philosophers and scientists have not yet agreed on whether there exists free will in humans. Both sides come with strong arguments. Determinists refer to causality as the basic principle ruling the existence while indeterminists claim that while causality is valid, the outcome cannot be predicted with absolute certainty. Some advocates of free will postulate that free will represents an original cause itself.

There have been various approaches by computer scientists aiming at reflecting volition or even free will in artificial intelligence. Approaches vary from creation of deterministic will that is called "free" to proposals to emulate free will resembling human mental functioning.

In this chapter at first a deterministic model will be described and assessed from the AGI's point of view. Next, an approach to human-like free will shall be presented. Finally, one of intermediate stages will be mentioned as well.

In 1988 John McCarthy proclaimed that with regard to free will "the robots we plan to build are entirely deterministic systems" [15]. Later in 2002 – 2005, he proposed a model of *Simple deterministic free will* [16] in which he reduced the notion of free will to (1) computing possible actions and their consequences, and (2) deciding about most preferable action. As an essential element he considers knowledge of choices. This approach refuses complexity of a system to exhibit free will.

Although this proposed model seems to be effective for the existing AI, it seems that such notion is not suitable for AGI purposes and emulation of human mental functioning since it is too simplistic. From psychological point of view, human mental processing is claimed to be based on three cooperating elements: volition, cognition and emotions [17]. Avoiding or reducing impacts of these elements in the processing then prevents making unpredictable solutions of which humans seem to be capable. Although some experiments have been made to disprove existence of free will (namely Libet's experiment), results of these experiments have been widely criticized and not fully accepted [8]. Moreover, unpredictability of human decisions is highly presumable with regard to the very biological nature of a human brain ("The brain is warm and wet, unpredictable, unstable, and inhomogeneous.") and principles of quantum physics [18]. According to those principles it is possible to determine only probabilities of future behavior but not exact future behavior [19].

An argument against existence of deterministic free will based on causality was also made by Perlovsky. He claims that causality reflected in logic is prominent in consciousness, but consciousness does not represent "a fundamental mechanism of mind" [13]. According to him in computer science dynamic logic is necessary to overcome the issue of complexity of mind that has own hierarchy. Moreover, conditions for existence of free will that can be formalized were already proposed and based on physical theories. These are said to be based on pairwise interactions of particles. Research shows that free will can in principle exist in case of interaction between three or more particles [19].

With regard to these facts it is obvious that a concept of free will should not be dismissed in AGI as inherently impossible or useless. It is, therefore, necessary to look at other, more complex models of free will emulation or their proposals. Much more favorable approach to artificial (mechanical) free will was taken by Manzotti. He claims that "free will stems out of very complex causal processes akin to those exploited by human beings. However, it is plain that simple deterministic devices are not up to the task" [20]. He states that freedom of an agent lies in capability of making real choices, i.e. choices that are not random but also not resulting only from external causes. He mentions a concept of gradual freedom in which freedom of an agent depends on its complexity and a degree to which individuality of an agent is expressed [20]. A degree of freedom in decision is also related to the degree of involved causal structures in an agent. An action resembling an instinct is considered to be much less free than an action involving own causal structure formed by individual history.

The presented technical approach is much more complex than simple deterministic free will. However, it does not provide any particular solutions. Only conceptual guidelines are outlined. Moreover, many constraints and problematic concepts to be solved are mentioned: temporal integration in an agent, polytropism, or automatic and conscious responses [20].

The two models, one of simple deterministic free will and the second of human-like free will, represent two ends on a scale of AGI development. It is obvious that any development is gradual (various approaches were briefly summarized by McCarthy and Hayes [21]); therefore, one needs to presume stages in which technology will improve over time. It has been shown that free will is rather difficult concept and includes many components. One of its main characteristics is unpredictability. As it has already been argued, the very unpredictability is caused by the biological structure of the brain [18]. Randomness represents its inherent feature. Therefore, this component should also be included in order to reach the next step in developing artificial free will.

In terms of computer science, however, creation of real random generator has been quite a difficult task to accomplish. The question is how can a deterministic model produce indeterministic results while it is working based on laws of logic? Software-generated randomness can be computed and is not then truly random. Fortunately, new research shows paths how to create real randomness. Some recent random generators are based on physical phenomena and use noise sources such as chaotic semiconductor lasers [22]. The most promising research is though in the area of quantum physics. Quantum randomness was proven incomputable and "is not exactly reproducible by any algorithm" [23]. The next step in developing artificial free will would then be incorporating and testing quantum random generator in order to provide AGI with a mechanism that can at any time provide it with a possibility to decide in a completely illogical way.

3.3 Possible Future Approach

Previous chapters outlined the current knowledge about volition, free will and attempts so reflect this characteristic in artificial intelligence. This last chapter should focus on the future and other possible steps or sources of inspiration for creating free will in AGI. One of the promising fields is neuroscience that studies neural systems and mechanisms that underline psychological functions.

With regard to biological basis of volition, very interesting research has been done by prof. Peter Ulric Tse who studied activity of neurons. Based on the results of his research he claims that free will has a neurological basis. According to his theory neurons react only in case some particular and predefined criteria are fulfilled. Decision of a person and her will are conditioned by the current structure and definitions. However, freedom of a person and her will lies in rapid neuronal plasticity. After a person made a decision, the neurons can reprogram themselves and define new criteria for future decision-making [24].

These findings are in fact in line with previous findings and specified psychological characteristics of free will. A person bases her decisions on previous experience. However, in case of employing complex cognitive processes, reaction can be changed for future cases. There is also delay in performing decisions by humans so it is presumable that before acting in a decided way, the particular person can quickly reconsider the action and act otherwise. To other humans such action seems instant and, therefore, free.

The comprehensive description of neural functioning by prof. Tse provides a great starting point for computer scientists to try to emulate similar functioning in the sphere of artificial intelligence. It seems to be the most feasible to use neural networks in order to achieve the same manner of functioning.

However, a serious limitation still persists even in this approach. Even when the activity of organic neurons would be perfectly emulated, it would be a mere presumption that as of this moment an artificial being has a free will. The problem with the free will is, as already mentioned, that this quality is dubious due to its first person perspective experience and cannot yet be even confirmed in animals. Further research in this field is necessary.

4 Conclusion

The aim of this paper was to frame the problem of free will in the context of AGI. In order to answer the question whether it is necessary to artificially construct free will to reach the ultimate goal of AGI two main problems were explored: (1) the nature and importance of free will for human mental functioning, (2) usefulness and technical possibility of its creation and incorporation into AGI.

It has been shown that free will as such significantly influences mental processing and overall behavior of a human. Characteristics associated with free will are considered to be uniquely human and contributing to development of intelligence.

In the field of AGI incorporation of such an element is presumed to bring significant improvement for agents situated in complex environments. Although there are many limitations and constraints yet to be solved, the possibility of creating free will seems to be viable and in case of continuous risk assessments also beneficial to the society.

The ultimate goal of AGI is to create a system that resembles or exceeds human capabilities in all areas including cognition and emotions. Since free will contributes to intelligence development, emotional control and possibly also self-awareness, and it seems to be construable, AGI needs to create this element to resemble human capabilities. Future attempts not only need to include real random generator that will be incorporated into the decision mechanism but also learn from neuroscience and get inspiration from mechanical functioning of the brain.

Last remark we wish to make concerns constructing a system that exceeds human capabilities. It needs to be noted that "exceeding human capabilities" is a very vague term. Since AGI aims at first to resemble a human, free will seems to be necessary. However, this will may also enable an AGI system to experience dilemmas, contradictions and human states in which it is sometimes difficult to make any decision. It is questionable which role free will plays in this drama. It can be at the same time the cause of all these problems as well as their solution.

References

1. Goertzel, B., Pennachin, C., eds.: Contemporary Approaches to Artificial General Intelligence. In: Artificial General Intelligence. Springer (2007)
2. Neisser, U., Boodoo, G., Bouchard Jr, T.J., Boykin, A.W., Brody, N., Ceci, S.J., Halpern, D.F., Loehlin, J.C., Perloff, R., Sternberg, R.J., et al.: Intelligence: Knowns and unknowns. American psychologist **51**(2) (1996) 77
3. Volition. `http://www.merriam-webster.com/dictionary/volition`
4. Kazdin, A.E., ed.: Volition. In: Encyclopedia of Psychology. Oxford University Press (2000)
5. Free will. `http://www.merriam-webster.com/dictionary/free%20will`
6. O'Connor, T.: Free will. In Zalta, E.N., ed.: The Stanford Encyclopedia of Philosophy. Spring 2013 edn. (2013)
7. Frith, C.: The psychology of volition. Experimental Brain Research **229**(3) (2013) 289–299
8. Haggard, P.: Human volition: towards a neuroscience of will. Nature Reviews Neuroscience **9**(12) (2008) 934–946
9. Baumeister, R.F., Muraven, M., Tice, D.M.: Ego depletion: A resource model of volition, self-regulation, and controlled processing. Social Cognition **18**(2) (2000) 130–150
10. Vohs, K.D., Schooler, J.W.: The value of believing in free will. encouraging a belief in determinism increases cheating. Psychological science **19**(1) (2008) 49–54

11. Ryan, R.M., Deci, E.L.: Self-regulation and the problem of human autonomy: Does psychology need choice, self-determination, and will? Journal of personality **74**(6) (2006) 1557–1586
12. Hong, F.T.: On microscopic irreversibility and non-deterministic chaos: Resolving the conflict between determinism and free will. In: Integral Biomathics. Springer (2012) 227–243
13. Perlovsky, L.: Free will and advances in cognitive science. ArXiv e-prints (2010)
14. Fisher, M.: A note on free will and artificial intelligence. Philosophia **13**(1-2) (1983) 75–80
15. McCarthy, J.: Mathematical logic in artificial intelligence. Daedalus **117**(1) (1988) 297–311
16. McCarthy, J.: Simple deterministic free will. `http://www-formal.stanford.edu/jmc/freewill2/`
17. Corno, L.: The best-laid plans modern conceptions of volition and educational research. Educational researcher **22**(2) (1993) 14–22
18. Donald, M.J.: Neural unpredictability, the interpretation of quantum theory, and the mind-body problem. arXiv preprint quant-ph/0208033 (2002)
19. Urenda, J.C., Kosheleva, O.: How to reconcile physical theories with the idea of free will: from analysis of a simple model to interval and fuzzy approaches. In: Fuzzy Systems, 2008. FUZZ-IEEE 2008.(IEEE World Congress on Computational Intelligence). IEEE International Conference, IEEE (2008) 1024–1029
20. Manzotti, R.: Machine free will: Is free will a necessary ingredient of machine consciousness? In: From Brains to Systems. Springer (2011) 181–191
21. McCarthy, J., Hayes, P.: Some philosophical problems from the standpoint of artificial intelligence. Stanford University (1968)
22. Kanter, I., Aviad, Y., Reidler, I., Cohen, E., Rosenbluh, M.: An optical ultrafast random bit generator. Nature Photonics **4**(1) (2009) 58–61
23. Calude, C.S., Dinneen, M.J., Dumitrescu, M., Svozil, K.: Experimental evidence of quantum randomness incomputability. Physical Review A **82** (2010) 022102
24. Tse, P.U.: The Neural Basis of Free Will. Criterial Causation. The MIT Press (2013)

Soft Competence Model for Case Based Reasoning

Abir Smiti and Zied Elouedi

LARODEC, University of Tunis
Institut Supérieur de Gestion de Tunis
Tunis, Tunisia
smiti.abir@gmail.com, zied.elouedi@gmx.fr

Abstract. Case based reasoning (CBR) is a methodology for building intelligent computer systems. An effective case base depends on axiomatic and key criterion: the case base competence which is measured by the range of problems that can be satisfactorily solved. Actually, modeling case-base competence is a delicate issue in the area of CBR. The proposed competence models in the literature are not robust enough towards noisy data. Furthermore, they have difficulties in handling the challenges posed by the collection of natural data which is often vague. To resolve these problems, we present, in this paper, a soft competence model, based on fuzzy clustering technique named Soft DBSCAN. Experiments are provided to show the effectiveness of our model and its high accuracy for predicting.

Keywords: case based reasoning, soft competence model, soft DBSCAN

1 Introduction

One of the great targets of Artificial Intelligence is to construct smart methods and systems able to recognize and follow human reasoning. Among these systems, Case Based Reasoning (CBR) [1] is a diversity of reasoning by analogy. CBR is able to find a solution to a problem by employing its luggage of knowledge or experiences which are presented in form of cases. To solve the problems, CBR system calls the past cases, it reminds to the similar situations already met. Then, it compares them with the current situation to build a new solution which, in turn, will be incorporated it into the existing case base (CB). Compared to other AI approaches, CBR allows curtailing the effort required for knowledge acquisition and representation significantly, which is unquestionably one of the supreme reasons for commercial victory of CBR applications. Actually, CBR has been used to invent innumerable applications in a spacious range of domains including medicine, games, management, financial, customer support, etc. Different ways have been recommended to illustrate the CBR process, but the traditional [1] the four REs CBR cycle (REtrieve, REuse, REvise, REtain), is the most frequently used: The new problem is matched against cases in the case base and one or more similar cases are retrieved. A solution suggested by the matching cases is then reused and tested for success. Unless the retrieved case is a close match, the solution will probably have to be revised producing a new case that can be retained.

In point of fact, the CBR donates better results as long as cases cover a large range of problems. Nevertheless, when the number of cases grows to a horrible high level, the quality of the system would downgrade. To cope with this problem and to guarantee the system's vigor, maintenance of CBR system becomes inescapable. Accordingly, in the CBR research branch, a great amount of concentration has been rewarded to Case Base Maintenance (CBM) [2, 3].

Actually, the potency of a CBM strategy depends on the quality of case data. Its performance can be measured according to a major criterion: Competence or named coverage which is the range of target problems that can be successfully solved [4]. Generally, the proposed approaches for case base coverage can be susceptible to the existence of unpleasant cases such as noises which are those whose descriptions are academic in nature and if learned in a case base, may cause the solutions to be bogus. Besides, many cases have approximately uniform coverage and others have very small coverage; thus, it is difficult to differentiate between these case's categories. Moreover, they have difficulties in handling the challenges posed by the collection of natural data which is often vague.

This new paper will fix attention to the important of soft computing in order to model the competence of the case base. The main idea is to repartition the case base into similar groups (competence groups) to ensure that the distribution of each group is nearly uniform, and more importantly the use of an efficient soft clustering technique named Soft DBSCAN to differentiate case's types and to compute the overall competence.

The advantage of our new proposed model is its elevated accuracy for predicting competence. In addition, it is not fragile to noisy cases, as well as it takes into account the vague data with uniform distribution. The extremely encouraging results obtained on some data sets are shown and discussed.

The rest of this paper is organized as follows. In the next Section, we introduce a quick overview of strategies for modeling the competence of the case, where we concentrate on the well known competence model proposed by McKenna & Smyth [5]. Section 3 describes our new soft competence model in detail and discusses their advantages. Experimental setup and results are given in Section 4. The paper concludes in Section 5.

2 Case-Base Competence: Related Work

Recent works [3, 6] highlighted the importance of the competence in a maintenance case base process. A case becomes useful when it improves the competence of the CBR system. We can define the competence, or coverage, by the set of problems that a case, or case-base, can solve. Nevertheless, it is tricky to quantify the competence of the system, for the reason that the accurate nature of the relationship between the case base and competence is multifaceted and not very well implicit. So, we require a theoretical model that permits the competence of a case-base to be estimated and guessed. Several proposed models to represent the coverage, in the literature [3, 7, 8] have demonstrated how dissimilar cases can create very different kinds of competence contribution.

The works proposed in [4] and [7] had highlighted the importance of determining the competence through adaptation costs and a similarity metric. They had delineated two key essential ideas which are coverage and reachability. In order to have a CB with excellent competence, its coverage ratio must be lofty.

Authors in [9] consider a case is momentous in the CB if it covers many related cases: its similarity value (sim) should be superior to a threshold Ψ. Based on many tests, the threshold Ψ can be defined using an hierarchical competence model.

The coverage model proposed by McKenna & Smyth [5] is a respectable involvement of the analysis of case base structure by evaluating the local competence contributions of cases and their relations. It is hypothesizing that the competence is based on a number of issues including the size and the density of cases. The number and density of cases can be readily gauged. In fact, the individual competence contribution of a single case within a dense collection will be inferior than the contribution of the same case within a sparse group; dense groups contain larger redundancy than sparse groups. The density of an individual case can be described as the average similarity between this case and other clusters of cases entitled competence groups. Therefore, the density of a cluster of cases is measured as an entire as the average local density over all cases in the group. The coverage of each competence group is an approximation of the problem space area that the group covers. As designated above, group coverage must be directly relative to the size of the group but inversely comparative to its density [7].

These works have drawn attention to the magnitude of modeling CBR competence. On the other hand, they endure from some inadequacies such as they do not take care about the situation of non-uniform distributed case-bases, as shown in [10]. Further, they scan the entire case base for the labeling of cases which is not obvious. In addition, they are touchy to mistaken cases like the noises. Besides, they do not handle the challenges posed by the collection of natural data which is often vague.

3 SCM: Soft Competence Model

To amend these troubles quoted above, we suggest, in this paper, a novel model for computing case base coverage, named SCM – Soft Competence model for Case based reasoning based on soft computing techniques. These include a new soft clustering technique named Soft DBSCAN and fuzzy Mahalanobis distance to differentiate case's types and to compute the overall competence.

As we have mentioned above, the determination of the case base competence is not a trivial process. The best way is to locate some ballpark figures to this set. Theoretically, we judge that the case base is a delegate illustration of the problem space. Under this condition and in order to smooth the progress of the competence computing, a specified case base can be crumbled into sets of similar cases. The competence of the case base as a whole is computed as the sum of these group coverage values. As they have been proved in [6], the competence of the case base is proportional to the individual coverage contribution of a single case within a resolute groups distribution, which is allied to the size of the case base. Hence, the total competence percentage of a given CB can be calculated as follows:

$$Comp\%(CB) = |1 - \frac{\sum_{j=1}^{k} \sum_{i=1}^{n} Cov(x_{ij})}{SizeCB}| \tag{1}$$

where k is the number of groups and Cov is the coverage contribution of each case in one cluster j with given distribution. This value depends on the type of the case and its role in the CB.

As a matter of fact, we have delineated three central types of cases which should be regarded as the key to achieve a good estimate of coverage computing:

- CN_i: Noisy cases are a distortion of a value or the addition of the spurious object. They are disagreeable cases, they can dramatically slow the classification accuracy. The best choice in this situation is to detect cases expected to be noisy and affect them an empty set as a coverage value ($Cov(CN_i) = \emptyset$).
- CS_i: Each case from a group of similar cases and which is near to the group centroid, provides similar coverage values, because they are close to each other, they cover the same set of cases. Hence, the coverage value of each case equals to the number of cases in this group (n). ($Cov(CS_i) = n$).
- CI_i: In a set of similar cases, there are cases that are much distant to the other members in this group. We can consider them as isolated cases. They belong to one set of similar cases not like those of type CN_i but they are farther to the set's centroid than the cases of type CS. They cover only themselves. As a result, the coverage of each case of this type equals to one. ($Cov(CI_i) = 1$).

Based on these explanations, we build a new coverage model named SCM – Soft Competence model for Case based reasoning based on soft clustering and a new efficient soft clustering technique. Our plan gets fix on these three sorts of cases and affects the suitable coverage value to each type.

To apply this idea, we necessitate first to craft multiple, small groups from the case base that are situated on different sites. Each small group holds cases that are closely related to each other. This can be done only by a clustering procedure because it guarantees that each group is small and surrounded similar cases, so it is effortless to spot the different types of cases.

After that, for each small cluster: the cases, which are near to the cluster's center and close to each other, are considered as cases of the type of CS_i. The cases, which are far away from the center, are considered as cases of the type of CI_i. Finally, the cases, which are outside the clusters and have not affected to a determined cluster, are considered as cases of the type of CN_i.

3.1 First Step: Clustering

Among the proposed clustering approaches, we should, ideally, use a method that while clustering and creating groups of similar cases, can smooth the discover of the different types of cases in such data sets. In particular, we want an algorithm that can manage instances expected to be noisy, it can create clusters with different shapes, and it allows the elements to have a degree of membership for each cluster. To overcome all these conditions, we use a new fuzzy clustering method named Soft DBSCAN proposed in [11]. We can resume the basic steps of our Soft DBSCAN as follows:

Algorithm 1 Basic Soft DBSCAN Algorithm

```
Begin
m: weighting exponent ($m > 2$)
$\xi$: tolerance level
Run DBSCAN and find:
        x = number of noises
        k = number of clusters
$c \leftarrow x + k$
Create the initial fuzzy partition:
        if $x_i \in c_j$ then $u_{ij} \leftarrow 1$
        Else $u_{ij} \leftarrow 0$
$t \leftarrow 0$
Repeat
       Update $U_t$ as following: cr$\mu_{ik} = [\sum_{j=1}^{c}(\frac{MD_{ik}}{MD_{jk}})^{\frac{2}{m-1}}]^{-1}$
       Where $MD_{ik}$ is the Mahalanobis distance between $x_k$ and $v_k$
        Calculate $v_t$ as following
       $v_i = \frac{1}{\sum_{k=1}^{n} \mu_{ik}^{m}} \sum_{k=1}^{n} \mu_{ik}^{m} x_{ik}$        i= 1, 2,...,c
Until $\|U_t - U_{t-1}\| \leq \xi$
$(U, v) \leftarrow (U_t, v_t)$
noisy points = $\{x_{ij} | cj = x_{ij}\}$
End
```

3.2 Second Step: Distinguishing the Different Types of Cases

Once we have partitioned the original case memory by Soft DB-SCAN, we select cases which are detected by our clustering technique. For these cases (CN_i), we accord them an empty set as coverage value.

Following, our soft competence model directs attention to finding the other types: As we have mentioned above, the CS_i are situated in the core of the cluster space, they pursue a standard distribution and they arise in a elevated possibility area of this cluster. On the other hand, the isolated cases CI_i are situated at the margin of the cluster space and diverge strongly from the cluster distribution. They have a squat probability to be produced by the overall distribution and they deviate more than the standard deviation from the mean.

For each cluster determined by the first step, each case gives weighted value of the cluster, and the weight is given by the membership degree of the fuzzy membership function. Heretofore, the cases which are distant from the core of the cluster are judged as cases of type "Isolated cases" CI. The paramount practice to spot them is the Mahalanobis distance with weighted mean and covariance of the cluster, and the weight is given by the membership degree of the fuzzy membership function.

We choose to use this distance because it takes into account the covariance among the variables in calculating distances. With this measure, the problems of scale and correlation inherent in the other distance such as Euclidean one are no longer an issue. In addition, Mahalanobis distance is an efficient for the non uniform distribution and arbitrarily shaped clusters because it deals with clusters of different densities and shapes.

$$MD_{x_i,V_i} = ((x_i - V_i)^T F_i^{-1}(x_i - V_i)^{1/2} \tag{2}$$

Where V_i gives weighted mean of the cluster, and the weight is given by the membership degree of the fuzzy membership function and F_n

is the fuzzy covariance matrix of the i-th cluster. Here, the covariance is weighted by the membership degree in the fuzzy membership function and defined by:

$$F_i = \frac{\sum_{k=1}^{n}(\mu_{ik})(x_{ik} - V_i)(x_{ik} - V_i)^T}{\sum_{k=1}^{n}(\mu_{ik})} \tag{3}$$

Where μ_{ik} is the membership degrees defined in the first step by the Soft DBSCAN clustering technique.

Based on our hypothesis, the cases with a large Mahalanobis distance in a cluster are selected as CI type. The brink of bulky distance depends on when the similarity between cases and the center starts raising. For that, we need to compare the MD of each case by the standard deviation of this cluster, in order to measure how closely the cases cluster around the mean and how are spread out in a distribution of the cluster. This last is able to know how firmly cases are clustered around the center. It indicates how much, on average, each of the cases in the distribution deviates from the center of the distribution because it depends on the average of the squared deviations from the mean of the cluster. Therefore, it is a good measure of the categorization of CI and CS cases, such the case whose MD is superior to the standard deviation of the cluster, will be consider as CI case, else it will be CS type.

As a result, we have affected for each case the appropriate coverage value depending on its type.

4 Experimental Analysis

In this section, we shall use experimental results to show the performance of our soft competence model SCM. We experimented with diverse data sets with different sizes and non-globular shapes, obtained from the U.C.I. repository [12] (Iris with size of 150 cases, Ecoli with 336 cases, Breast W with 699 case, Blood T with 586 cases, Indian with 786 cases and Mammographic with the number of 961 cases). Our target is to show that the model proposed in this paper offers a good correlation to the case base accuracy.

In the first part of our experimentation, our competence model was applied to each case-base and its predicted competence compared to the test set accuracy, where we apply the 1-NN algorithm to the same datasets and the same task to obtain the average accuracy rate. Initially, the training set was partitioned into five randomly independent sets. We use, in this situation, the correlation coefficient, in order to measure the relationship between the CB accuracy and the predicted competence model. Actually, this coefficient is a number between 0 and 1. If there is no relationship between the predicted values (our SCM competence) and the actual values (PCC), the correlation coefficient is 0 or very low. Figs. 1 shows the experimental corollaries of three different datasets which have been presented in different five case bases.

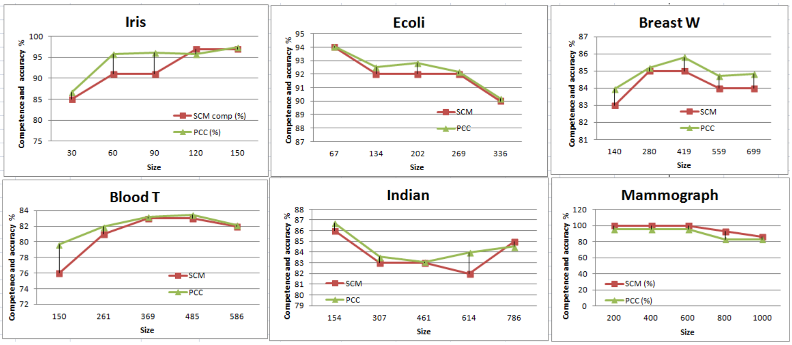

Fig. 1. Comparing predicted competence (SCM) to the case-base accuracy (PCC) for different Case-base Sizes.

The results give commendable efforts in support of our soft competence model. It comes into views to be a very adjoining correlation between the two curves and hence a strong relationship between forecasted competence granted by our SCM model and the test set accuracy.

5 Conclusions

The study of this paper has drawn interest to a chief measurement in the Case Based Reasoning, which is case base competence or case base coverage. In this paper, we have proposed a novel soft competence model for case based reasoning. It is based on an efficient soft clustering technique named "Soft Dbscan" and fuzzy Mahalanobis distance to differentiate case's types and to compute the overall competence. The benefit of our suggested model is the employ of soft techniques in order to exploit a tolerance for imprecision, as well its soaring accuracy for foretelling CBR's coverage. Besides, it marks the character of the distribution and the special types of cases such as the noises, isolated cases and similar cases. The results of experiments shepherded are very influential and optimistic for our model. Future tasks include applying this model in the maintenance of the case based reasoning systems.

References

1. Aamodt, A., Plaza, E.: Case-based reasoning: Foundational issues, methodological variations, and system approaches. Artificial Intelligence Communications **7**(1) (1994) 39–52
2. Leake, D.B., Wilson, D.C.: Maintaining case-based reasoners: Dimensions and directions. Computational Intelligence **17** (2001) 196–213
3. Smiti, A., Elouedi, Z.: Overview of maintenance for case based reasoning systems. International Journal of Computer Applications **32**(2) (October 2011) 49–56 Published by Foundation of Computer Science, New York, USA.
4. Smyth, B., Keane, M.T.: Remembering to forget: A competence-preserving case deletion policy for case-based reasoning systems. In: Proceedings of the 14th International Joint Conference on Artificial Intelligent. (1995) 377–382

5. Smyth, B., McKenna, E.: Competence models and the maintenance problem. Computational Intelligence **17**(2) (2001) 235–249
6. Smiti, A., Elouedi, Z.: Modeling competence for case based reasoning systems using clustering. The 26th International FLAIRS Conference, The Florida Artificial Intelligence Research Society, USA (2013) 399–404
7. Smyth, B., McKenna, E.: Building compact competent case-bases. In: Proceedings of the Third International Conference on Case-Based Reasoning, Springer (1999) 329–342
8. Reinartz, T., Iglezakis, I., Roth-Berghofer, T.: On quality measures for case base maintenance. In: Proceedings of the 5th European Workshop on Case Based Reasoning, Springer Verlag (2000) 247–259
9. Grachten, M., García, F.A., Arcos, J.L.: Navigating through case base competence. Proceedings of the 6th international conference on Case-Based Reasoning Research and Development (2005) 282–295
10. Shiu, C., Yan, L., Wang, X.Z.: Using fuzzy integral to model case-base competence. In: Workshop on Soft Computing in Case Based Reasoning. (2001) 301–310
11. Smiti, A., Elouedi, Z.: Soft dbscan: Improving dbscan clustering method using fuzzy set theory. In: 6th International Conference on Human System Interaction, HSI, IEEE. (2013)
12. Asuncion, A., Newman, D.: UCI machine learning repository. http://www.ics.uci.edu/mlearn (2007)

In Search for Memory: Remembering from the Viewpoints of Philosophy and of Multi-agent Systems Research

Ondřej Beran

Research Center for Theory and History of Science,
Faculty of Philosophy and Arts, University of West Bohemia,
Plzen, Czech Republic
ondrej.beran@flu.cas.cz

Abstract. The paper deals with the concept of memory that is subject to philosophical discussions. Using the development of multi-agent systems as a comparative example, I try to show that the popular storage concept of memory is incorrect. It is more appropriate to consider memory as a strongly associative capacity mixing procedural with propositional elements (among others) and emerging from a network of collective (intersubjective) interrelationships, than as a (spatial) property of a specified part or location of either organic body or a machine. This intuition is more clearly promoted by the development at the field of computer science and engineering than by debates in philosophy.

Keywords: memory, storage concept, multi-agent systems

1 Introduction

In this contribution, I deal with the philosophical discussion about memory and the viewpoint from which it can be illuminated in a useful way by the advances in creating an effective artificial memory. The outline of the paper is as follows: in the first part, I present perhaps the most popular folk psychological notion of memory, the so-called storage concept, and some of its philosophical advocates. In the second part, I introduce some serious philosophical and scientific arguments in favor or a more complex view on memory. In the third part, I sketch briefly the (from a considerable part historical) place of the storage concept in the context of the computer memory. In the fourth part, I present, in a very short glimpse, the advances in the computer memory facilitated by the development of multi-agent networks. In the final part, I suggest that the computer engineering advances are illuminating for philosophical discussions of what memory is and how it works because they embody, unintentionally, some of the boldest and most controversial philosophical insights, particularly concerning the collective nature of memory, and – due to the technology-specific and practice-oriented nature of the branch – because they avoid some typical philosophical shortcomings of metaphysical provenience.

2 Storage Concept of Memory

In *A Study in Scarlet* by Sir Arthur Conan Doyle, the great detective Sherlock Holmes surprises his friend, Doctor Watson, by his complete ignorance of the heavenly mechanics. But not only he does not know whether the Earth runs around the Sun or the other way round, his lack of interest in such topics and its explanation provided by Holmes is equally curious:

> Now that I do know it [= the real movings of the heavenly bodies – O.B.] I shall do my best to forget it. ⋯You see, I consider that a man's brain originally is like a little empty attic, and you have to stock it with such furniture as you choose. A fool takes in all the lumber of every sort that he comes across, so that the knowledge which might be useful to him gets crowded out, or at best is jumbled up with a lot

> of other things so that he has a difficulty in laying his hands upon it. Now the skilful workman is very careful indeed as to what he takes into his brain-attic. He will have nothing but the tools which may help him in doing his work, but of these he has a large assortment, and all in the most perfect order. It is a mistake to think that that little room has elastic walls and can distend to any extent. Depend upon it there comes a time when for every addition of knowledge you forget something that you knew before. It is of the highest importance, therefore, not to have useless facts elbowing out the useful ones. [1]

Holmes summarizes here what we could call the storage concept of memory, characteristic for many folk psychological intuitions: memory is a place with certain defined (determinate) storage capacity. It can diminish or extend within time, but in a particular time moment t the capacity of the memory storage is given. Memories kept within memory are data units of a fixed, specific volume. Only certain bulk of memories can be contained in t and if the capacity has been filled, nothing more can come in (unless, perhaps, something that has been in comes out).

Apart from the folk psychological intuitions, there are also respectable philosophical champions of the storage concept, mostly from the representationalist tradition of Modern philosophy (such as Descartes, Locke or Hume). Later, Bertrand Russell offered a distinct representational theory of memory, incorporating also *propositional* elements. Unlike the Modern tradition according to which the storage contains *image-shaped* ideas, Russell points that what is retrieved from the storage has propositional content linked to the agent's various other knowledge (the strength of this interlinking distinguishes memory from, say, imagination) [2]. Among recent authors continuing this vein Jerry Fodor can be named, for whom representations of events/experiences are stored within memory in a form translatable to our language [3, 4]. Non-philosophical theorists expressing a storage view on memory (e.g., psychologists) are more difficult to find, yet there are some (for survey see [5]). One respectable advocate is Endel Tulving [6, 7], despite his reserves.

In some respects, there are good reasons why the storage concept is quite popular. We commonly observe and say that only some bulk of material "fits into" one's memory (or more literally: one's head), and if there is too much inside, no new material can come in, or only at the expense of loss of the previously stored material. As among scholarly theorists, so among folk psychologists and everyday speakers there is no agreement whether memories are stored in an isomorphic shape, strongly resembling the remembered, or in some less straightforward form, as well as no agreement as to the supposed causal (neural?) background. The shared point is that memory is a space – literal or figural – into which only some bulk can fit and be stored for a time.

3 Other Perspectives of Memory

The storage concept of memory is not the only one possible and has been criticized extensively. Memory is not necessarily a thing; it is *what happens* when we remember something and "there are many very different processes of remembering", as Wittgenstein [8] puts it. Psychologists distinguish among several types of what we call memory, among which the storage concept occupies only a limited position. There is the "procedural memory" of a rather bodily nature, by virtue of which we are able to perform various practices we have adopted. The procedural memory enables us to display our knowledge-*how*, or more simply: skills. The nature of this memory is kinaesthetic and there is no remembered (that is: stored) content to identify.

Memory can also have a propositional form in which we evoke a knowledge-*that*: a memory either of situations from our lives (episodic memory) or of various factual contents (who killed Caesar and in what year). The propositional form tempts to searching for the evoked content, its form and possibly also its storage place. It is, however, quite difficult to identify any such (causal?) link between memory expressed and memory stored; meaningful memories exist, as such, rather in a rich conceptual context linking them to other memories, knowledges etc. [9, 10]. The propositional memory has a basic claim to be true and without being true it is actually useless for anything else [11]; but there is no reasonable way of establishing truth of a proposition with reference to an uncertain (internal) material of a non-propositional nature.

Wittgenstein [12] criticizes the attempts to trace anything causal (stored) behind our capacity to remember things; though he admits he knows nothing about the possible memory background. But for our understanding of what memory is such search is useless. The brain activity behind memory needn't take shape of specified stored contents or representations of anything. Memory is a capacity of an agent (a being, an individual), not a property of a specific part of the body (say, brain) – a proper understanding requires avoiding this "mereological fallacy" [13].

Among various scholarly theorists a dynamic, operational view on memory is more common, according to which memory is something that is being performed and in this sense it does not exist in a "stored" form during the time when it is not performed [14]. Naturally, the propositional content of memory as something performed in real time by bodily, socialized agents facilitates the view on memory as something constructed with inclusion of various pragmatic and contextual coefficient factors [15, 16].

Against the folk psychological storage conception also dissenting folk psychological intuitions can be used. The storage concept systematically underestimates the *associative* nature of memory. Associative capacity enables elaborating and developing the memories into unexpected details and consequences; in a sense, the more there is in memory, the more can be "stored" there in addition. On the other hand, such a memory has often a "fractal" nature, as we going into more and more details of the already remembered; whereas the storage concept assumes the ability to keep a specific bulk of memory contents *regardless* of their dis-/continuity. Anecdotic observations could confirm both these intuitions: in practice, the capacity-limited memory keeping discontinuous contents acts along with the associative memory capable of elaborating the remembered content into greater and greater width.

4 Representations: Talking about Computer Memory

Can machines have memory? From one point of view, the question is pointless, given the uncontested fact of the computer memory. Yet Wittgenstein, e.g., questions this trivial fact, pointing that it is only of certain type of agents (living human beings typically) that *we say* that they are conscious, feel pain, or also have memory [17]; if memory is understood as propositional, then it *tells* (not: shows) us something about the past, and memory put in such terms can be conceived meaningfully only with respect to beings capable of telling or understanding the told [12]. Our present, 70 years later habit of talking (meaningfully) about the computer memory, including all the ramifications of the subject shows that the conceptual patterns have evolved radically.

The question of memory in computers was, in its beginnings, linked rather closely to the storage view. The pioneer of AI debate Alan Turing [18–20] suggested the computer memory as a – potentially – limitless reservoir (in contrast to the human memory which is limited). In this reservoir, data and instructions are stored and can be retrieved from it, provided that we have a proper reading routine capable of searching among the stored symbols. This Turing's idea is interesting in its difference from the human memory in two respects: i) it presupposes a quite open and strict storage concept, while the human memory is a rather complex conjunction of capacities; and ii) unlike the storage concept embedded in most folk psychological intuitions about the human memory, it postulates an infinite storage capacity to be brought about someday.

This conception of memory in computers, initiated by Turing, was rather typical in his age; and it is still a default outsider understanding of what "memory" in computer context means. When philosophers criticize certain types of accounts of memory – especially the representationalist ones – they argue, in this simplistic vein, that it is perhaps the computer memory what works the way representationalists portray the human memory, but not the real human memory [21].

It cannot be overlooked that in computers, the storage devices enabling that there is a memory at all are indispensable. Designers have had to deal with this from the very beginning as with the crucial technical problem; a storage system had to be designed and constructed before there could be any memory. (At any rate, memory is not quite the same as storage in computers either; memory is a software process running on a hardware background and using the storage capacities; this seems neglected by the critics of the computer-like accounts of the human memory.)

The problem for philosophers may lie in the very representationalist point of the metaphor. If memory is a representation of something, it can be more or less true (true or false). In the case of computers, the truth of the stored material is the central condition: their memory presupposes that the information material can be retrieved in exactly the same form it was stored (memorized, so to speak). The experience of the users of the common PCs is analogous: memory is the capacity of the computer to store some information material and, on demand, to reproduce it again in identical form. Many (but not all the) data stored in computer have been put in, intentionally, by the end-users of the PC for whose needs (and from the point of view of whom) the stored-retrieved identity is realized.

In this respect, the problematic analogy even with the storage part of the human memory seems to collapse. The translation between the remembered event and memory in the propositional form is not easy; there can be two adequate memories of the same event put into very different words. What is more important, in humans there is no guarantee (of the kind of the inputting user of PC) that the stored content stays the same in the meantime; one can never be quite sure whether what she/he remembers is "the same" all the time or whether it has altered with respect to the remembered fact [21]. The design and the construction of the "storage device", if there is any such thing, is also unknown. While in humans the storage concept seems to be simply incorrect (not just insufficient), it provides a solid description of the computer memory or at least of how "folk computer science" – from some good reasons – understands the working of PCs.

5 Beyond the Storage Concept: Multi-agent Networks

Computer engineers, however, didn't content themselves with considering computers as passive machines that can only render, exactly in the same form, the data stored in them by their human "masters". The storage parallel in people allows double reading of the storage limit: either one just cannot store more than certain bulk in one's head, or she/he is just not able unlimitedly to recall discontinuous information. In the second reading, it is the agent's ability to remember what is concerned, along with the question of *how to make someone/something remember*, disregarding memory as representation storage. The second perspective on memory is also more favored in the recent software engineering as the more fruitful and interesting in practice.

A breakthrough in the computer memory was conferred by multi-agent systems. What acts as an agent is not necessarily a computer unit in the strict physical sense, but rather an autonomous computer *system* situated in an environment. These agents have to be equipped with decision-making systems (architectures) enabling them to act within certain range of situations (to deal with certain range of problems). From this point of view, control systems such as thermostats or software demons are also agents, autonomous to certain extent. Agents are connected to one other in communication networks through which they not only exchange information but perform what can be considered as analogues of (human) social interactions and act as self-interested [22]. The cooperation among agents does not require that each agent follows instructions defining the whole work or task; the information and instructions are distributed among them and no one is in charge of the whole. The performance is, however, modeled as a performance of a *group* [23].

Already in its beginnings, the multi-agent systems research emphasized the agents' capacity to learn and display what they have learnt. Learning is not founded within hard-wired agents, but it is the online interaction among individual agents what is crucial. The collaboration enables agents to improve in problem-solving procedures and recall the efficient moves for further application [24]; for which the memory capacity to make "indexes of experiences" [20] is useful. Their utility consists not in retaining a bulk of specific data, but in facilitating the effective orientation within – either old or new – available information with respect to problem-solving. The information landscape displayed in the corporate memory of a multi-agent network is not structured homogeneously or arbitrarily, but it is clustered by the preferences of the web users. Since the users of the web are themselves clustered into various "organizational entities" (professional, political, hobby-related, etc.), the structure of the information landscape parallels the meaningful structures of the organization of human institutions [25]. The "natural", embodied associativity of the human memory growing from the individual's situatedness in a rich social and institutional context is substituted by such preference-shaped architecture of the information landscape accessed by the multi-agent network. The individual agent is then both a member of a network and a situated and limited unit with a unique pattern of heterogeneous access to different parts of the information landscape; just as individual humans differ from one other in the composition of their memories and information they can recall; quickly or after some time, directly or through a means or another agent.

It was argued that individual agents are self-interested – each agent has its own description of the preferred state and strives for reaching and keeping it. This prescription needn't involve any (negative) instruction towards other agents or (positive) towards itself, but comes in terms of problem-solving [23]. The agents also strive for increasing the "basin of attraction" for their preferred states by which they modify the structure of connections in the whole network. Some say [26] that the individual agents display an analogue of Hebbian [27] learning: they act selfishly, without any ambition to create an associative memory. Nonetheless, by such self-organizing they alter the dynamics of the whole system in a way that creates an associative memory of a system (the ability to recall past patterns of activation under similar initial conditions). A thereby originating memory transcends the simple propositional memory and shifts towards the procedural memory: a distributed multi-agent system exhibit adaptive behavior analogous to human capacity to learn. This cognitive capacity of a system emerges as a function from the moves of individual agents that bear no ambition to improve the capacity of the system [26].

What is interesting about this approach to memory is that there is no one bearer of the memory, easy to identify. The historical efforts at creating the computer memory focused on the hardware level; in the multi-agent network research, memory is not something that is stored (that is, permanently located) somewhere as an informational content representing something outward. The persistence of an agent in terms of the persistence of what is within the reach of its memory then becomes, naturally, quite a different (perhaps far more difficult) enterprise than just keeping a physical storage unit at work [28]. Memory is a *capacity*, but not even as a capacity it can be attributed to one bearer, be it a physical computer unit or a computer system. Storage can be physically built, memory cannot. Memory emerges from a network.

6 Impulses for Philosophy

The subtleties of multi-agent system designing are indeed interesting, but AI theorists and professionals needn't be lectured about them from a philosopher who lacks the proper expertise. In the preceding section, I summarized briefly some of the main points from the field of multi-agent research. Actually, rather than considering their real respective importance, I wanted to point what stroke a philosopher as interesting for a philosophical treatise of memory. An account of memory *per se*, so to speak (it is no surprise that by "memory" philosophers regularly, implicitly mean the human memory) can learn a good deal from the engineering approach.

The strategy adopted by computer scientists – to focus on the level of networks of agents who needn't be of strictly speaking physical nature – is interesting and instructive for philosophy. Unlike the quarreling philosophers, the former effectively avoids the reifying *metaphysics* of memory that would identify memory with a thing or a location that can be pointed to with a finger or even taken by a hand. Memory is thus no longer dealt with as storage. But not because there are final, irrefutable arguments; there will always be also arguments supporting that in some contexts, it makes solid sense to speak of memory as of some kind of storage facility (for philosophers at least). The reason is that in the *practice* of computer engineering there are much more fruitful ways how to model memory. This way, the meaning of the word "memory" and the way we approach memory-endowed entities as such is established through and in situated practice, from below, independently of an explicit theory of memory.

The practice also implicitly accepts that the notion of memory is mixed in its nature; meaning that it contaminates procedural and propositional elements. (We may constantly overestimate the importance of memory as propositional: demonstrating a memorized information – "Caesar was killed in 44 BC" – can be interpreted as a skill, too, while there is nothing propositional about a soldier demonstrating that she/he, after some time has elapsed since her/

his training, still remembers how to dismantle a gun with her/his eyes blindfolded. Whether this skill consists of a representation of some kind is unclear.) From this point of view, it would be futile to focus only on the question of storing the remembered representations somewhere inside the memory-endowed agents. It took considerable effort on the side of some philosophers – such as Wittgenstein – to explain that we don't need to trace the representations stored behind the once-present experiences or knowledge, or that there are very many things called "memory" or "remembering", and that in many cases, it makes no good sense to speak of retrieving a stored content at all (typically in the cases of the procedural memory). Yet there is no easy agreement on the field of philosophy and the champions of the representationalist storage concepts such as Fodor still influence the debate considerably.

Therefore, the focus put *not* on mentalist metaphysics but on integral and autonomous agents and their performances, interconnections and networks remains an epistemically profitable trend primarily – if not only – in the context of computer science. In philosophy, the interest in the internal build-up of memory-endowed agents and the interest in agents' memory performances in practice still wait for reconciliation. Post-analytic philosophers instructed by Wittgenstein or Davidson (who endeavored at turning the explanation of beliefs, skills etc. from the inward to the outward view) can only appreciate this "perspicuous representation": a machine that can be taken by hand (a PC unit, e.g.) does not carry the attributes of something we consider as thinking or endowed with a full-fledged memory. The entities endowed with such a memory – the whole multi-agent networks or agents as members in a network – do not have any clearly identified "inside" where their memory could be searched for as a property of a piece of anything.

This is a link to another reason why philosophers concerned with memory should pay more attention to the achievements on the computer field: the elucidating way of demonstrating memory to be a *collective* matter. This line of argument is usually very indirect in philosophy: one has to demonstrate i) that human memory performances emerge from a background of conceptual intuitions, practical habits etc. ii) and that these cannot originate at all – or at least in their actual shape – without the agent being involved in a rich network of intersubjective relationships. (Both steps – and there being two steps only is a simplification – are no minor tasks for a philosopher.) With respect to the artificial memory we see rather straightforwardly that the present achievements couldn't have been realized without the multi-agent environment. Facilitating such a collective environment is a *necessary* prerequisite for reaching an effective memory which includes elements constitutive also of the adult human memory: such as associativity or procedural, performative (problem-solving) memory skills. On the other hand, no strong metaphysics of collectivity is evoked: individual agents require no "awareness" of their being part of a precisely defined group, their relations towards the others needn't be specified. The performing collective is not sustained by a leading idea or a leading agent.

Some recent research in AI, e.g., in evolutionary robotics contributed interestingly to our knowledge of the evolution of thought [29]. What computer engineering can tell us philosophers about memory – what is memory and how it originates and works – is equally interesting, I think. But the possibilities of such an interdisciplinary exchange with respect to memory have not been explored yet to a comparable width.

Acknowledgments. This study was supported within the project of Education for Competitiveness Operational Programme (OPVK), Research Centre for Theory and History of Science (Výzkumné centrum pro teorii a dějiny vědy), registration No. CZ. 1.07/2.3.00/20.0138, co-financed by the European Social Fund and the state budget of the Czech Republic.

I would also like to thank Radek Schuster; his useful discussion hints are partly responsible for encouraging me to write this text.

References

1. Doyle, A.C.: A Study in Scarlet. Penguin Books (2011)
2. Russell, B.: The Analysis of Mind. G. Allen & Unwin (1922)
3. Fodor, J.A.: The Language of Thought. Harvard University Press (1975)
4. Fodor, J.A.: LOT 2: The Language of Thought Revisited. Oxford University Press (2010)
5. Roediger, H.L.: Memory metaphors in cognitive psychology. Memory & Cognition **8**(3) (1980) 231–246
6. Tulving, E.: Elements of Episodic Memory. Clarendon Press Oxford (1983)
7. Tulving, E.: Episodic memory and autonoesis: Uniquely human? In: The Missing Link in Cognition: Origins of Self-Reflective Consciousness. Oxford University Press (2005) 3–56
8. Wittgenstein, L., Rhees, R.: Philosophische Grammatik. Volume 5. Suhrkamp (1973)
9. Campbell, J.: Past, Space, and Self. MIT Press (1995)
10. Campbell, J.: The structure of time in autobiographical memory. European Journal of Philosophy **5**(2) (1997) 105–118
11. Poole, R.: Memory, history and the claims of the past. Memory Studies **1**(2) (2008) 149–166
12. Wittgenstein, L.: Bemerkungen über die Philosophie der Psychologie. University of Chicago Press (1980)
13. Moyal-Sharrock, D.: Wittgenstein and the memory debate. New Ideas in Psychology **27**(2) (2009) 213–227
14. Toth, J.P., Hunt, R.R.: Not one versus many, but zero versus any: Structure and function in the context of the multiple memory systems debate. In: Memory: Systems, Process, or Function? Oxford University Press (1999) 232–272
15. Barnier, A.J., Sutton, J., Harris, C.B., Wilson, R.A.: A conceptual and empirical framework for the social distribution of cognition: The case of memory. Cognitive Systems Research **9**(1) (2008) 33–51
16. Campbell, S.: Relational Remembering: Rethinking the Memory Wars. Rowman & Littlefield (2003)
17. Wittgenstein, L.: Philosophical Investigations. John Wiley & Sons (2010)

18. Turing, A.M.: On computable numbers, with an application to the entscheidungsproblem. In: The Essential Turing. Clarendon Press (2004) 58–90
19. Turing, A.M.: Intelligent machinery. In: The Essential Turing. Clarendon Press (2004) 410–432
20. Turing, A.M.: Intelligent machinery, a heretical theory. In: The Essential Turing. Clarendon Press (2004) 472–475
21. Stern, D.G.: Models of memory: Wittgenstein and cognitive science. Philosophical psychology **4**(2) (1991) 203–218
22. Wooldridge, M.: An Introduction to Multiagent Systems. Wiley (2008)
23. Shoham, Y., Leyton-Brown, K.: Multiagent Systems: Algorithmic, Game-theoretic, and Logical Foundations. Cambridge University Press (2009)
24. Garland, A., Alterman, R.: Multiagent learning through collective memory. In: Adaptation, Coevolution and Learning in Multiagent Systems: Papers from the 1996 AAAI Spring Symposium. (1996) 33–38
25. Gandon, F., Dieng-Kuntz, R., Corby, O., Giboin, A.: Semantic web and multi-agents approach to corporate memory management. In: Intelligent Information Processing, Kluwer Academic Publishers (2002) 103–115
26. Watson, R.A., Mills, R., Buckley, C.L.: Global adaptation in networks of selfish components: Emergent associative memory at the system scale. Artificial Life **17**(3) (2011) 147–166
27. Hebb, D.O.: The Organization of Behavior: A Neuropsychological Approach. John Wiley & Sons (1949)
28. Barthès, J.P.: Persistent memory in multi-agent systems – the omas approach. International Journal of Energy, Information and Communications **2**(4) (2011) 1–14
29. Rohde, M.: Enaction, Embodiment, Evolutionary Robotics: Simulation Models for a Post-Cognitivist Science of Mind. Springer (2009)

Beyond AI: Artificial Golem Intelligence
Proceedings of the International Conference Beyond AI 2013

Editors:
Jan Romportl, Pavel Ircing, Eva Zackova, Michal Polak, Radek Schuster

Typesetting: Pavel Ircing
Cover design: Tom Vild
Printed: Typos, tiskařské závody, s.r.o.

Published by University of West Bohemia, Pilsen, 2013
First edition

ISBN 978-80-261-0275-5