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

Beyond AI: Artificial Dreams

Proceedings of the International Conference
Beyond AI 2012
Pilsen, Czech Republic, November 5th–6th, 2012

ISBN 978-80-261-0102-4

Published by University of West Bohemia

Preface

If you read this page, you probably already have an idea of what this book is about. To see whether your idea is right, you should probably start to read its chapters (or at least table of contents) because this preface can provide you merely with somehow empty phrases such as "interdisciplinary and philosophical aspects of artificial intelligence", "deep-rooted ideas of AI", or "controversies of AI". And this preface indeed does not want to provide you with any empty phrases.

This preface wants to be a way in which I can express my deep gratitude to everyone who made the conference "Beyond AI: Artificial Dreams" possible. It means that my great thanks go to a number of people who most probably already know it even though their names are not listed here. Yet there are couple of people who were not only incredibly helpful in all organisational issues, but without whose support on so many different levels the conference would not happen at all: Eva Žáčková, Radek Schuster, Pavel Ircing and Michal Polák.

Pilsen, October 2012

Jan Romportl
Organising Committee Chair
BAI 2012

Organisation

Beyond AI: Artificial Dreams (BAI 2012) 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 and Department of Philosophy of the same university. The conference took place in Pilsen, on November 5th and 6th, 2012.

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)
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 (Loquendo, Torino)

Organising Committee

Jan Romportl
Eva Žáčková
Pavel Ircing
Radek Schuster
Michal Polák

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.

Keynote Talks

Heteronomous Humans and Autonomous Artifacts: The Paradox of AI
Hamid Ekbia (Indiana University)

Going to the Dark Side—the Misuses of AI Promises
in Persuasive Health Technologies
Søren Holm (University of Manchester)

Is a Singularitarian Ethics Impossible?
David Roden (Open University)

Moral AI: What Kind of Morality Should AI Have?
Julian Savulescu (University of Oxford)

The Disappearing Human-Machine Divide
Kevin Warwick (University of Reading)

Table of Contents

Humean Machine: When Desires Are in Charge

Ivo Pezlar

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

Abstract. It is already broadly agreed upon, although not yet deeply studied, that motivational relations such as desires and goals should play important role in devising artificial agents. In this paper we propose different approach to modeling desire: instead of embedding desire into the reasoning mechanism, we insert reason into the desiring mechanism. In addition new distinction between desires and dummy-desires is introduced.

Keywords: desires, dummy-desires, motivational attitudes, formal models of desire, humean machine

1 Introduction

AI was born as goal-oriented, problem-solving discipline and having a goal alone was seen as sufficient reason for performing an action. In other words, goals themselves were seen not only as a cause, but also as a purpose of certain action: no difference was perceived between having a goal and desiring a goal.

For many tasks this simplifying approach to problem-solving works just fine. It would be hardly useful to have autonomously desiring toasters toasting only when they find it appropriate to do so. But tasks like toasting are relatively straightforward and what's more important we have already figured out how to do them. However, there are many other and much more complicated problems we don't know how to solve yet and ideally we would like AI to help us and assist us in coming up with the solutions.

I don't think I'm saying something wildly controversial or novel when I say that what is essential (at least for humans) for formulating a winning strategy (i.e., devising a plan that will achieve desired goal) is a desire to win. Put differently, first we have to want to find a solution to find a solution.

At first, it might not seem so significant (things get sometimes discovered by "accident", but then again it is usually during some other problem-solving

process), but you can hardly come up with a solution to a problem you don't want to solve at all.

But this all puts machines in awkward positions: on one hand we want them to solve problems as we do, but on the other hand we strip them from our main aid in problem-solving, i.e., the desire to see it done or solved.

It's not uncommon that we view our desires rather as weaknesses or obstacles during the problem-solving process – as something which rather stands in our way then helps ("I want to go out and have some fun with friends, I don't want to sit home alone trying to prove this theorem any more.") – but what if it is the desire that is essential in creative and autonomous problem-solving as hinted before? Put another way, by depriving machines of desires, we might willingly cripple their problem-solving capabilities.

To sum it up, we put forward the following idea that having a certain goal is not enough for effective formulation of successful strategy: it is also crucial that the agent (human or artificial) *wants* to achieve that goal.[1] Simply put, we see desire as a key trigger for problem-solving process.[2]

2 Desires and Dummy-Desires

What has been said so far is nothing really new under the sun. Desire and its influence on our actions is topic as old as philosophy itself and even taking desires into consideration in AI has been done before. There is of course the well-known and studied BDI theory [1] and its extended reincarnations (e.g., BDICTL [2], LORA [3] and many others), but they all perpetuate certain viewpoint that might limit creating strong AI.

We are talking about desires being handled just as any other part of the reasoning process – being inserted into the reasoning procedure – while in reality it seems rather the other way around. In other words, many try to subsume desires into the reasoning mechanism, even though they seem to be stand-alone: we don't reason with our desires; desires control, oversee,

[1] I'm deliberately avoiding here the question of machine consciousness, because I'm not convinced that it is necessary for desiring. Or at least not in the way that this paper is concerned. Of course, in the end it might turn out that consciousness is indeed essential for the study of desire, but until then I think there is no need to willingly limit our options with this yet to be resolved premise.

[2] In this paper we will not be distinguishing between desires and motivations: when we are desiring something we are also motivated to achieve it and and vice versa. Of course, this supposition is far from being unproblematic, but it's not significant for our discussion here.

supervise and set to motion the whole reasoning process.[3] And this is the position to which we will subscribe here.

Acceptance of this thesis leads to re-evaluation of BDI and similar theories: what they reason with cannot be desires any more, but rather their mere representations, which we may call *dummy-desires*.

By dummy-desires we mean pieces of information that share the content with "genuine" desires, but are lacking "the pull" that defines true desires, i.e., the drive that motivates us towards action. These pseudo desires carry over the information of what is desired, but lack the ability to trigger the action towards the desired. This power is reserved for true desires only (and which are in return inert towards the whole reasoning process which employs dummy-desires).

So what do we mean by true desires? True desire are simply those desires that cannot be reasoned with and which are capable of producing action. To put it differently, whether something is to be considered as a true desire depends solely on its detachment from reasoning process and simultaneously on its ability to trigger an action.[4]

This new distinction between desires and dummy-desires can explain quite easily why is it possible to desire premises of some argument, without desiring their logical consequence. The answer is: because what comes into play in reasoning process are not really desires, but just their "stunt doubles" for reasoning. Whatever happens to them does not effect the original desires upon which they were shaped. In other words, we use dummy-desires to test outcomes of our real desires without actually committing to them.

For example: let's say I posses desires p and $\neg q$ and I want to check their consistency (desire r). After some time I come to the conclusion that $p \rightarrow q$. Now it seems that I have three options: discard desire p, $\neg q$ or desire r. That's the rational conclusion. But the fact is I don't have to discard any of those. I might simply ignore the whole argument and keep on desiring p, $\neg q$ and r. Why? Because the rational argument didn't have any action-driving weight behind it. And why is that? Because what was used was not really desires, but their representations for reasoning which I call dummy-desires. Of course, you could accuse me of being inconsistent in my desires (and rightly so!), but that won't change anything about them.

[3] Recall Plato's well-known allusion in Phaedrus with chariot as a reason and two "desire" horses pulling it forward, i.e., putting it into motion.

[4] We are being here purposefully vague about the nature of desires themselves, so that we don't have to commit to some particular philosophical or psychological conception.

In other words, we cannot reason with our desires (e.g., I can desire not to reason at all). But what we can do is to reason with their "empty", hollow representations (i.e., dummy-desires), which we use in our arguments. E.g., if it is 2 AM and I have a craving for a steak, I might come to the conclusion that it is not the best idea to eat one pound of grilled meat just before sleep. So I decide not to eat it. But that doesn't mean I got rid of the desire to eat a steak. I just acted on stronger desire "to act rationally" or "to live healthy" etc.

But do we really need this distinction? Couldn't be the examples above explained e.g., in terms of competing and conflicting desires: in the sense that my desire to be healthy triumphs over my desire for a steak, therefore I choose not to eat a steak? Not really, because once we put this dilemma into this argument form it's obvious that I might still opt to ignore it and eat that steak after all.

The contrast between desires and dummy-desires cannot be translated into matter of competing desires. The same goes for reducing it into other distinctions such as e.g., first-order and second-order desires, goals and subgoals, etc. Dummy-desires are not weaker versions of desires; dummy-desires are rather "names" for real desires which are then used in reasoning.

To put it differently, the distinction between desires and dummy-desires is not one of intensity (stronger/weaker, higher/lower, long-term/short-term...), but of quality: the former has different roles and properties than the latter.

The key thing to keep in mind is that the main purpose of this distinction is to help us explain our somewhat paradoxical reasoning behaviour: on one hand we are definitely reasoning with our desires, but on the other hand we don't necessarily feel any pull to obey the outcomes.

To summarize: we do not reason with our desires (e.g., as BDI does), rather desires control our reason. Or as Hume [4] puts it:

> Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them. (...) Since reason alone can never produce any action, or give rise to volition, I infer, that the same faculty is as incapable of preventing volition, or of disputing the preference with any passion or emotion.[5]

So desires are not slaves of the reason (or in case of BDI of planning and believing), but more plausibly the opposite is true. So maybe we should no try to incorporate desires into reasoning process, but rather develop desiring process and try to include reasoning process in it.

[5] See Hume, D.: A Treatise of Human Nature. pp. 414–415. (Book II, Section III).

The main idea can be in a nutshell captured in what we shall call *humean machine*, i.e., machine where desires have the sovereignty over reason (Fig. 1):

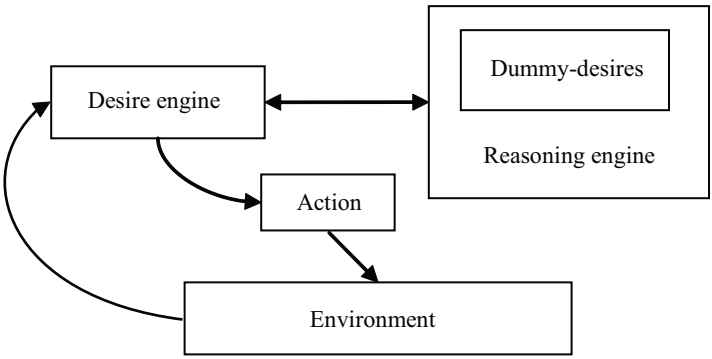

Fig. 1. Scheme of humean machine

It is important to note that the desire engine should have a "kill switch" over the reasoning process, so it can jump over it right to performing an action. In other words, there needs to be a desire to act rationally to employ the reasoning engine in the first place. Shortly put, it is the desire engine alone that triggers the reasoning engine.

Once we do this "humean turn", i.e., once we let desires control reason, it should be obvious that there has to be two way connection between desire engine and reasoning engine: we have (most of the time) desire to act rationally and be considered rational and in order to do that we first have to know *what* is the rational thing to do. And to find out that is the role of the reasoning engine.

It is this bidirectional connection that enables us to take into account our desires while reasoning. But it is important to remember that what we are reasoning with are not strictly speaking desires, but only their reflections, i.e., dummy-desires, which cannot make us move, to put it bluntly.

This desire/dummy-desire dichotomy helps us to explain why we don't always act rationally (or even morally for that matter). More specifically, why cogent and rational arguments, while we fully acknowledge their validity, do not need to have any persuasive power over us (even if we add clauses like "I want to act rationally"), e.g., there is nothing wrong with the following

statement: "I know I should do it and it's the rational thing to do and I want to be rational, but I still don't want to do it." That's all because what we are reasoning with are not our actual desires, but only their representations for reasoning, i.e., dummy-desires. Generally put, no rational argument alone could ever trigger an action on its own accord.

But how should we represent these desires and dummy-desires and what should be the inner workings of such a desire engine? In last section I try to address these questions.

3 Meet A.L.I.C.E.

First of all, we introduce our humean machine *Autonomously Longing & Intelligently Computing Engine* or A.L.I.C.E. for short. For representation of desires we will use modal propositional logic K,[6] which means that we will treat desire similarly as epistemic logic treats belief ([5], [6]). In other words, we assume that desires have propositional content.

We will read the modal formula $\Box p$ as "agent desires p" or "agent desires that p". Formally, the language L of modal desire logic is non-empty, countable set of atomic formulae Atom $= \{p, q, ...\}$ and formulae defined as follows:

- $\phi := p \quad | \quad \neg\phi \quad | \quad \phi \vee \psi \quad | \quad \phi \wedge \psi \quad | \quad \phi \rightarrow \psi \quad | \quad \phi \leftrightarrow \psi \quad | \quad \triangle\phi$

where symbol $\triangle$ will be used instead of $\Box$ to emphasize the fact that we are dealing here with desire, not necessity (or knowledge).

Next we add to any full axiomatization of propositional logic the following:

- Axiom 1: $\triangle\phi \rightarrow \phi$
- Axiom 2: $\triangle(\phi \rightarrow \psi) \rightarrow (\triangle\phi \rightarrow \triangle\psi)$
- Inference rule 1: from ϕ and $\phi \rightarrow \psi$ infer ψ
- Inference rule 2: from ϕ infer $\triangle\phi$

Semantics: structure M for language L is tuple $\langle W, R, V \rangle$ such that

- W is nonempty set of truth worlds,
- $R \subseteq W \times W$

[6] It's important to note that the manner in which we choose to represent desires (and dummy-desires) in A.L.I.C.E. is not really a key issue here and our choice of modal logic was motivated mainly by its simplicity and familiarity, which makes it well-suited system for basic exposition of the topic and also very solid starting point for further discussion (e.g., which axioms are most appropriate and so on).

- $V : W \times Atom \rightarrow \{\texttt{Goal}, \texttt{Not-Goal}\}$ is distribution function on atomic formulas,

Satisfaction relation $\models$ is defined as follows:

- $(M, w) \models p$ iff $V(w)(p) = \texttt{Goal}$
- $(M, w) \models \neg\phi$ iff $(M, w) \not\models \phi$
- $(M, w) \models \phi \wedge \psi$ iff $(M, w) \models \phi$ and $(M, w) \models \psi$
- $(M, w) \models \triangle\phi$ iff $(M, w_2) \models \phi$ for all $w_2 \in W$ such that $(w, w_2) \in R$

So we say that agent desires ϕ (written $\triangle\phi$) if and only if ϕ is her goal in all worlds she considers as possible (assuming her knowledge base stays the same across all the worlds).

The rationale behind this formalization is that desire drives us towards certain goals no matter what are the circumstances: if we want something, we want it even if it is impossible (or highly improbable). Of course, our desires might change accordingly to what we know, which is captured by the requirement of constant knowledge base throughout the worlds.

It should be obvious that we are talking here about very strong desires. E.g., I desire drinking water (i.e., agent desires such a state of affairs in which she is drinking water) if and only if I can't imagine world (with what I now know) in which it is not my goal to drink water.

Notice that so far we have taken into account only true desires, while omitting dummy-desires entirely. To amend this we need to introduce another modal system (basically duplicate of the one we have already introduced) and add metalanguage labels to its formulae to distinguish it from the earlier system for true desires. So in the end we would have p for desires (domain of desire engine) and p' for dummy-desires (domain of reasoning engine), where "$'$" is the metalanguage label.

The idea is that even if a conclusion of certain argument is for us to want p' (e.g., "I desire to eat a steak" is in our belief/knowledge base), we end up doing p' if and only if we have also p in our desire base.

So aside from desiring engine (producing p) and reasoning engine (producing p') we also need third component which would check if we really want to do what our reason tells us to do. In this respect, desires might be considered subordinated to even higher reasoning: some sort of meta-reasoning which compares matches of desires with dummy-desires, but this description would be slightly misleading, because our third component is not so much engaged in reasoning as in evaluating and checking desires and dummy-desires. So more fitting name would be meta-desiring, i.e., desiring that we act upon desires that are rational.

This concept allows us in very rudimentary manner model desire-dependent reasoning, i.e., we will do what we think is rational (p') if and only if there is match between p and p', i.e., if we also want to do p in the first place.

Of course, most of the questions are still ahead. What needs to be done next is e.g., devising the mechanism for converting the full-blooded desires into dummy-desires and then sending them to the reasoning engine. This is only the sketch at best. However, our task is somewhat simplified by the fact that that much of the research on desires has been already done (e.g., within the scope of decision theories). We just have to examine those theories and look for something that might be fruitful and easy to implement for AI purposes: we don't have to start from scratch.

Final summary: We focus too much on the "problem-solving" aspect of AI, while neglecting the desires driving the problem-solving itself. We need systematic, formal and rigorous account of internal motivation which has fully under control the reasoning mechanism. McCarthy and Hayes were back in 1969 hypothesizing and "organ of will" [7], now it might be the right time to do the very same with "organ of desire".

And just to reiterate and perhaps clarify, our goal is to allow machines to have the same starting position as we seemingly have during problem-solving procedure, i.e., having reasoning process that is governed by desires. Of course, the desires themselves can (and should) be "punched in" by designer. That is not what is principal here, what is important is the way in which the machine will work these desires (pre-set or not). What is then the relation between reason and desire? Reasoning helps us to (re)evaluate our desires (to see what is rational and what is not) and thus influence our actions and behaviour (if we desire to be rational), but what it does not do is directly controlling our actions.

References

1. Bratman, M.E.: Intention, Plans, and Practical Reason. Harvard University Press (1987)
2. Rao, A.S., Georgeff, M.P.: Modeling Rational Agents Within a BDI-architecture. In: Allen, J., Fikes, R., Sandewall, E. (eds.) Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning, pp. 473–484. Morgan Kaufmann Publishers Inc., San Mateo, CA, USA (1991)
3. Wooldridge, M.: Reasoning about Rational Agents. Intelligent Robotics And Autonomous Agents. MIT Press (2000)

4. Hume, D.: A Treatise of Human Nature. Clarendon Press (1896)
5. Hintikka, J.: Knowledge and Belief: An Introduction to the Logic of the Two Notions. Contemporary Philosophy. Cornell University Press (1962)
6. Fagin, R., Halpern, J., Moses, Y., Vardi, M.: Reasoning About Knowledge. MIT Press (1995)
7. McCarthy, J., Hayes, P.J.: Some Philosophical Problems from the Standpoint of Artificial Intelligence. In: Machine Intelligence, pp. 463–502. Edinburgh University Press (1969)

From Gobble to Zen
The Quest for Truly Intelligent Software and the Monte Carlo Revolution in Go

Ralf Funke

SNAP Innovation, Hamburg, Germany
ralf.funke@snap.de

Abstract. After the success of chess programming, culminating in Deep Blue, many game programers and advocates of Artificial Intelligence thought that the Asian game of Go would provide a new fruitful field for research. It seemed that the game was too complex to be mastered with anything but new methods mimicking human intelligence. In the end, though, a breakthrough came from applying statistical methods.

Keywords: Go game, Monte Carlo, UCT, chess

1 The Turk

In 1769 an amazing machine was introduced to the world, the Automaton Chess Player, known to us now as the Mechanical Turk. For more than eighty years the Turk was exhibited all over Europe and in the United States and showed his ability to play chess – winning most of his games. He is said to have played against Frederick the Great, Napoleon and Benjamin Franklin.

In retrospect it is hard to believe that the Turk could have been taken seriously at all. After all how could one imagine a machine being constructed that was able to recognize a chess position, to move the chess figures and to win against even quite strong players at a time when the most advanced technological breakthrough was the mechanical clock and certain music automatons. It would take nearly two hundred years more and the industrial and computer revolution to have some real artificial chess playing devices.

But although some people suspected a hoax from the beginning, it seems that many, if not most of the people, believed that a chess playing automaton was possible. In 1836 Edgar Allen Poe tried to explain the "modus operandi" of the Turk in an essay called *Maelzel's Chess-Player*. He states that one could

find "men of mechanical genius, of great general acuteness, and discrimitive understanding, who make no scruple in announcing the Automaton a pure machine, unconnected with human agency. . . " [1]

Well before the advent of Artificial Intelligence the history of the Turk teaches an important lesson. People are likely to exaggerate the ability of their engineers and maybe to underestimate the complexity of certain human endeavors.

Poe, after mentioning a couple of real automatons, like the famous duck of Vaucanson, goes on to compare the Turk with the calculating machine of Charles Babbage. He rightly claims that a chess playing machine, were it real, would be far superior to a calculator since "arithmetical or algebraic calculations are, from their very nature, fixed and determinate. . . " And so the results "have dependence upon nothing [...] but the data originally given". In chess, on the contrary, no move follows necessarily from the previous. After a few moves no step is certain. And even granted, Poe says, that the moves of the automaton were in themselves determinate they would be interrupted and disarranged by the indeterminate will of its antagonist. He continues with some technical objections to the mechanical Turk and then adds a very strange argument: "The Automaton does not always win the game. Were the machine a pure machine this would not be the case – it would always win." The difficulty of constructing a machine that wins all games is not "in the least degree greater [. . .] than that of making it beat a single game". This might be dubbed the Poe fallacy.

If the willingness of 18th century people to believe in the possibility of highly complex automatons is somewhat surprising, it should be remembered that the belief in a purely mechanistic and thus deterministic universe dates back at least another 150 years to the work of Galileo and to that of William Harvey, who following Fabricius, discovered blood circulation and showed that the heart was just a pumping machine and to Descartes who was prepared to announce that all animals were in fact automatons. Descartes, it has been argued, was influenced by the technological wonder of his time, the Royal Gardens created by the Francini Brothers, with their hydraulic mechanical organ and mechanical singing birds [2].

In the dualistic tradition it is the hallmark of the human agent to act in a non-determinate way, thus creating a new branch in the tree of life. This ability was what Poe denied the Automaton.

When the first computers were developed it seemed logical to create chess playing programs. A program to beat an expert human player would surely have capacities that would go far beyond arithmetical calculations. It would

need to have what would later be called Artificial Intelligence. It would need to be able to make choices based on the evaluation of a complex position.

2 The Game of Go

The story of the development of chess playing programs is well known. From the humble beginning of Turing's theoretical considerations to Deep Blue it took less than 50 years.

Creating a program that is able to perform at world championship-level is surely an astonishing accomplishment, but at the same time there are grave doubts whether one could call a chess program in any sense intelligent.

Of course, one could judge the performance simply by the results, and then the program must be regarded as intelligent or more intelligent than the players it beats. And it was known by Turing that any goal in computer science that is reached would be declared trivial afterwards, followed by the examples of feats that computers will never be able to accomplish. But still, the suspicion that high class chess programs are basically only sophisticated number crunchers, not principally different from the calculating machine of Babbage, remains a strong one.

No one really knows exactly how human players judge positions and what processes go on that result in the decision to play one particular move, but it is surely totally different from the way the computer works. And, if truly intelligent behavior is defined as behavior similar to that of humans, chess programs are not intelligent.

Maybe then, chess is just not complex enough, to really require true intelligence. Fortunately, there is one game that had the reputation of being so deep that it could never be played successfully by game tree analysis, the game of Go.

> This has given rise to the intriguing notion that Go is in fact the classical AI problem that chess turned out not to be, that solving Go will in fact require approaches which successfully emulate fundamental processes of the human mind, and the development of these approaches may both give us new insight in to human thought processes and lead to the discovery of new algorithms applicable to problems ranging far beyond Go itself. [3]

And indeed it has been said that Go has become the most exciting challenge for AI and can be regarded the final frontier of computer game research [4]. What is it then that makes Go special? Go, like chess, is a two

person, zero-sum, complete information game. But the board is larger and a typical game takes about 250 moves (in Go a move is a ply, or what is a half-move in chess).

The number of possible positions in chess are 10^{43}, in Go about 10^{170}. The whole game complexity can be calculated to be 10^{67} in chess compared to 10^{575} in Go [5].

The number of possible games is not the main issue though, since even on small boards, (9 x 9 is customary for beginners, humans as well as programs), the game remains complex. The reason is that there is no simple evaluation of a board position. In chess it is possible to weigh each figure on the board and together with some relatively simple heuristic rules (a knight at the edge of the board is worth less than in the centre) one can get a fairly accurate value of the position. In Go on the other hand it is sometimes not easy to decide whether a move increases the value of a position for one side and very hard to compare the relative virtues of two candidate moves.

3 The Rules

The rules of Go are very simple.

Preliminary Rule: Go is played on a 19 x 19 board with black and white stones. One player called Black takes the black stones one player called White takes the white stones. Black starts and then both players play alternate moves until both players agree that the game is over.

Principal rule of Go: A move can be played on any empty intersection of the board (including edge and corner) and remains on the board unless all adjacent points are taken by the opposite stone color.

Exception of the rule: A stone may not be placed on an intersection, if all adjacent points are taken by the opposite color. (Suicide Rule)

Exception of the exception: A stone may be placed on an intersection that is completely surrounded by enemy stones if the empty intersection point is the last empty adjacent point of this enemy stone – or a chain of enemy stones, where a chain is defined as stones of one color where every stone has at least one adjacent neighboring stone. (Capture Rule)

Exception of the exception of the exception: A stone may be not be placed on an empty intersection, even if this takes the last free adjacent point of one enemy stone, if the stone that would be so captured has itself captured exactly one stone with the previous move. (Ko rule)

Secondary Rule: The advantage of having the first move is compensated by a certain number of points (*Komi*) given to White. Large differences in

strength are compensated by a number of so called handicap stoned that are placed at the beginning of the game on the board.

The object of the game is to put as many stones on the board as possible.

This is not the set of rules that you would find in Go books. In the real world there are Japanese and Chinese rules (and even New Zealand rules) that differ slightly and add certain nuances. Especially the last point, the object of the game, would normally be defined in a different way. The object really is to surround as many empty points and capture as many enemy stones as possible and the game ends when no meaningful moves are possible.

But implementing this set of rules is enough to create a Go-playing program.

For a human player learning these rules is not nearly enough to understand the essence of the game. In practice, a novice at the beginning very often learns certain concepts that really follow from the rules. Especially important is the concept of a living group. A group lives, i.e. can never be captured, if it has two eyes, he will learn. An eye is a point surrounded by neighboring stones of one color. (The concept of a living group follows from the suicide rule.) But sometimes a group can have a false eye and then it is dead. And really a group does not need to have two eyes, it just must have the potential to build two eyes, if necessary, i.e. when it is attacked. Sometimes a group looks alive but is really dead, because within the group there is a "dead shape" of enemy stones. And what exactly is a group? A group is a collection of single stones or small chains positioned roughly in the same area of the board, in other words what constitutes a group is a fuzzy concept. Only when it is really alive, it is clear which stones belong to the group. So, the player decides what to regard as a group. He has to decide if a group is dead or alive, if it is weak or strong, if it can be connected to some other group or if it has to live internally. The player must learn to appraise the status of his own groups, but at the same time that of his opponent. And in the end he even has to learn how and when to sacrifice a group. The player will learn to play "good shape" moves and to avoid bad shapes. He will probably learn a couple of hundred defined sequences in the corner (called *josekis*), sequences that are regarded to give an equal result to both players, and any number of "proverbs" like "death lies in the *hane*". He will learn the sometimes very subtle difference between a forcing move that strengthens the own position or creates some potential and a move that really only strengthens the opponent. And very importantly, he will have to learn the value of keeping the initiative, of leaving a local fight to play somewhere else first. This is known in Go as keeping *sente*, as opposed to *gote*.

It seems clear that a Go playing program must have access to the kind of knowledge described here in one form or another. Some aspects of go knowledge are easy to implement. A program can reference a database with corner sequences to pick a *joseki* move. The same is true for good and bad shape moves. In a local fight the correct sequence of moves to kill an enemy group or to make life for an attacked group might be reached by brute force tree search. But some of the other concepts, like evaluating the status of a group or when to switch to a different part of the board are notoriously hard to put into code.

The attempt to establish "expert systems" was made all the more difficult as a lot of knowledge is implicit and cannot easily be put into words much less into code. For example the Go proverb "Play the important move first, then the big one" is often repeated but hard to appreciate.

There have been a number of different approaches to create a Go playing program [4], [5]. In theory the best idea seems to be to just implement the basic rules and let the program learn everything on its own. Some attempts have been made in this direction but they did not go very far.

In practice, it seemed, that "Go programmers must observe human Go players and mimic them." [6] And in the end it came down to the problem of how a move is to be evaluated. To judge the merits of a move there seem to be only two ways, namely a direct evaluation based on heuristics or a full board static evaluation after the move.

Direct evaluation is sometimes possible, e.g. when a move makes life for a big group. And sometimes one can hear commentaries such as: "White has played only good moves, black on the other hand has played one dubious move, therefore the position must be better for white." But certainly every amateur player knows from experience the situation, where he thinks that he has made the better overall moves, and still his position is worse than that of the opponent.

Because a full tree search is practically impossible in Go it was a natural idea, to regard Go as a sum of local games. In a local situation it is much easier to find a good or even the best move. And this is how a human player behaves. He will very often concentrate on one or two local positions, pick a couple of candidate moves in that position "that suggest themselves", and then try to falsify them. In the end the move is played for which the player could not find strong answers for his opponent. But in the context of game programming, this introduces a new problem. Even if a local perfect move is found, then the resulting local position has to be compared to other local positions. For example, it might be possible that there are two moves, both ensuring life to two different groups in jeopardy, then it might be the case

that it is better to save the smaller group, if this group plays an active role in the game and the other is of no strategic value. Of course this is only a sub problem resulting from the main problem that no fast and reliable full static evaluation of a board position was known.

It is no surprise then, that progress in computer Go was slow. At the end of the 90s the best Go programs were said to be around 3rd kyu, which would have been respectable if true. A beginner starts roughly as a 35th kyu and as he gets stronger the kyu grade steps down until first kyu is reached. Then the next step is first dan and then the dan grading climbs up. Very strong amateurs are 5th or 6th dan. The 3rd kyu rating was mainly for marketing purposes. In a very famous game, played in 1998, Martin Müller played a 29 stones handicap game to one of the strongest programs at the time, "Many Faces of Go", and won. (The game can be found in [4].) This would make the program roughly 25th kyu or really just the strength of a beginner. Müller is a Go programmer himself and knows the weaknesses of programs, but even taken this into consideration, programs could not have been much stronger as 10th kyu then. A fresh idea was needed to take computer Go forward.

4 Monte Carlo

In 1993 Bernd Brügmann presented a program called "Gobble" that introduced a new principle to the world of Go programming that would eventually trigger the Monte Carlo revolution of Go [7]. Monte Carlo techniques had been used before in physics or in mathematics, for example to solve the travelling salesman problem for practical purposes.

Applied to Go the basic idea is, that candidate moves are evaluated by starting simulated games from the current position with this move and to play random moves from there on, till the end of the game. For every considered move hundreds and now many thousand random games per second are played and the average score of the playouts is assigned to the move. Instead of taking the actual result only the win or loss is counted in most Monte Carlo implementations these days.

If this leads to good results, this approach has two obvious advantages to the standard way of Go programming. It practically needs no Go knowledge and since the counting at the end of game is trivial, it eliminates the need to evaluate a current position. The only real Go knowledge needed, is that the program needs to know that in playing the random games one should not fill one's own eyes. But it would be very easy to add a rule that forbids such virtual suicide.

Brügmann admitted that the idea might appear ridiculous. But he showed that in his implementation Gobble could play at 25th kyu on a 9 x 9 board, which was very impressive for a program without knowledge. And even if it is hard to accept that random moves could give an indication of a good actual move to play, it does make sense that starting random games with the first move in the centre of a 9x9 board leads more often to a win, than starting somewhere on the first line.

It did take a couple of years for the idea to really ignite. Ten years later Bouzy and Helmstetter take up the idea and add some refinements [8]. For one thing Brügmann had used not only the result of games that started at a particular move but also the value of the move if it was used in other simulations provided it was played the first time. The rationale for this was the observation that some moves are good no matter when they are played. Also, the moves played in a random game were not completely random but played with a probability that was dependent of their current value. This was to ensure that good moves had a better chance of being played. And some algorithm also controlled the probability that a move could be played out of order.

The value of the all-moves-at-first-heuristic was questioned and instead progressive pruning was introduced, where a move after a minimal 100 random games would be pruned, if it was inferior to another move. What is important though, is that the modifications were all in the realm of statistics.

It would take another statistical algorithm, though to help the Monte Carlo method in Go to its breakthrough. In 2006 the UCT algorithm was suggested for Go playing programs [9]. UCT means Upper Confidence Bounds applied to Trees. UCB was first used to solve the so called multiarmed bandit problem. It means that a formula is used that will guarantee that a move chosen for sampling will be either one that has already a good value and looks promising or a move that has not been sufficiently explored. This "exploitation vs. exploration" principle was used in the program "Mogo", which won the 2007 Computer Go Olympiad and was the first program to beat a human professional player at 9 x 9 Go [10]. Today all leading Go programs use the Monte Carlo/UCT method. The best probably being "Zen" which has reached a 6th dan rating at 19 x 19 on the popular KGS Go Server.

Some other improvements of statistical evaluation have been added like RAVE (Rapid Action Value Estimation), which allows to share information between similar positions (it is related to Brügmann's all moves as first heuristic) and some caching techniques. And, of course, based on the solid Monte Carlo platform even some Go knowledge is now used to prune or bias moves.

Even Many Faces of Go has reached 2nd dan, combining now its traditional Go knowledge with the Monte Carlo Tree Search.

Within six years, since 2006, the situation has changed dramatically. Before then every moderately serious Go player, say half of all club players, could beat any Go program without difficulty. Today maybe less than 1 percent of all amateur players can beat the strongest Go programs. This is the result of the Monte Carlo revolution in Go.

5 Conclusion

From the viewpoint of Artificial Intelligence the success of the recent development in Go programming obviously, and maybe sadly, repeats the history of the research in chess programming. In fact the way strong Go programs work now, does not even remotely resemble an emulation of "fundamental processes of the human mind". A chess program does what a human brain can at least aim at: consider as many follow up moves as possible to a candidate move and then evaluate the resulting position. Nothing like this could be said for Monte Carlo Go.

Bruno Bouzy who had spent many years developing a Go program, "Indigo", with standard Go heuristics and was then one of the godfathers of Monte Carlo Go summarizes and ends his activity with this remark:

> In 2006 and 2007, with the birth of the Monte-Carlo Tree Search technique, computer go is now in the right direction, like computer Chess was with alfa-beta. The future improvements in computer go depend on parallelisation and UCT refinements. The way from knowledge to Monte-Carlo is succeeded. Consequently, I suspend Indigo development for an undetermined period. [11]

This may be a bit of an overstatement since Go knowledge does play a role, but one can sympathize with his attitude.

If Go like chess failed to meet the expectations of Artificial Intelligence it might be a good idea to define intelligence other than in reference to a human being.

One of the pioneers of computer Go, Allan Scarff, came up with this definition:

> The degree of scope for appropriate behavior of an agent for any given set of knowledge and any given amount of processing used by that agent. [12]

The less knowledge is needed the more intelligent an agent is. In this respect Go programs are doing fine, but of course they need a lot of "processing", which according to this definition is a mark of the unintelligent.

José Capablanca, the chess champion, is supposed to have answered the question how many moves he would look ahead thus: "Only one, but it's always the right one." A program will never accomplish this, but then Capaplanca's mastery in chess was certainly the result of a lot of work and acquired knowledge. And just because a lot of the "pruning" and "biasing" happens unconsciously, it does not mean that not a lot of processing of some kind is going on.

And even if the best Go programs today can beat strong amateurs, there is still a long way to go to reach the level of top professional Go players. It may very well be the case that Monte Carlo Go leads to a dead end. Perhaps entirely new concepts have to be developed to really master the game. It might be the case that the human way is after all the most effective. But, I at least rather doubt it.

For one thing, intelligence is not the only aspect that is needed to reach top level, and maybe not even the most important. It is no coincidence that practically all professional players learnt the game in very early youth, and most did little else than studying Go. In this respect they resemble prodigies of, for example, piano playing. One of the best Go books is called *Lessons in the Fundamentals of Go* by Toshiro Kageyama. It is the grasping of fundamentals, Kageyama says and demonstrates, that differentiates the professional from the amateur (not only in Go). But the ability to grasp fundamentals, in contrast to appreciating them intellectually is something that is very hard if not impossible for an adult. And the reason is that active intelligence and a conscious desire to understand is an obstacle to absorb certain concepts. The human way for top achievements in Go, as well as in the arts, in sports, and the sciences is a very subtle interaction between rock solid fundamental knowledge outsourced into the realms of the unconscious and intelligent, creative, conscious application of this knowledge to specific circumstances.

This does not mean that it is the best way. The way human beings think and act is not something that is in principle denied to artificial beings. It might be possible to emulate the working relationship between consciousness and subconsciousness, and this would be very instructive, but I do not think that it is necessary in order to create artificial solutions for any task that seems at this moment to be restricted to the problem solving power of a human being.

To 19th century people it seemed that a machine, by definition, could not create something new, since it lacked free will and could only do what was

"built in". Today, it is not easy for a programmer, to even understand the problem. Any complex program will act in unforeseeable ways. This happens because of bugs, but just as easily by design if some random "decisions" are implemented. And in the same way as the program can act, as if it were free, it will act as if intelligent. For practical purposes there is no difference.

It might still be worthwhile to try to emulate human thinking, but there is no doubt that, as long as the quest for truly intelligent software comes up with highly original unexpected pseudo solutions like Monte Carlo Tree Search, we should not give up the quest.

References

1. Poe, E. A.: Maelzel's Chess-Player, http://www.gutenberg.org/files/2150/2150-h/2150-h.htm
2. Jaynes, J.: The Problem of Animate Motion in the Seventeenth Century. In: Kuijsten, M. (ed.) The Julian Jaynes Collection, pp. 69–84. Julian Jaynes Society (2012)
3. Myers, R. T., Chatterjee, S.: Science, Culture, and the Game of Go, www.bob.myers.name/pub/go-overview.doc
4. Müller, M.: Computer Go. Artificial Intelligence 134(12), 145–179 (2002)
5. Bouzy, B., Cazenave, T.: Computer Go: an AI-oriented Survey. Artificial Intelligence 132(1), 39–103 (2001)
6. Bouzy, B.: Spatial Reasoning in the Game of Go. http://www.math-info.univ-paris5.fr/~bouzy/publications/SRGo.article.pdf
7. Brügmann, B.: Monte Carlo Go. Technical report. Physics Department, Syracuse University (1993)
8. Bouzy, B., Helmstetter, B.: Monte-Carlo Go Developments. In: van den Herik, H. J., Iida, H., Heinz, E. A. (eds.) Advances in Computer Games 10: Many Games, Many Challenges, pp. 159–174 (2003)
9. Kocsis, L., Szepesvári, C.: Bandit Based Monte-Carlo Planning. In: Fürnkranz, J. et al. (eds.) Machine Learning: ECML 2006, LNAI, vol. 4212, pp. 282–293, Springer, Heidelberg (2006)
10. Gelly, S., Silver, D.: Achieving Master Level Play in 9 x 9 Computer Go. In: AAAI 2008, pp. 1537–1540 (2008)
11. http://www.math-info.univ-paris5.fr/~bouzy/IndigoHistory.html
12. Scarff, A.: AANNS explained. http://www.britgo.org/files/scarff/AANNS_Explained.pdf

A Novel Emotion-Based Decision Making Model to Use in Lifelike Intelligent Agents

Mohammadreza Alidoust

Islamic Azad University - Gonabad Branch, Gonabad, Iran
m.alidoust@hotmail.com

Abstract. Modeling behavior of intelligent agents and its affecting parameters is a very challenging aspect of research in the field of Artificial Intelligence. But, if performed correctly, we can improve the abilities of artificial agents and we can build social agents which can speak, think and behave like us. Many other models of behavior for intelligent agents have been proposed but their complexity makes it difficult to validate them against the real human decisions. In this paper a novel behavioral model is proposed which has a simple structure and also includes the effect of emotions as a major affecting parameter to the decision making process.

Keywords: intelligent agent, behavior modeling, decision making, emotion modeling

1 Introduction

Behavior modeling for intelligent agents is a new research aspect in control, computer science, sociology, psychiatry, psychology, economy, military, etc. The vast application field of this aspect varies from human-like robots, pet robots, and human behavior simulations in severe situations to building intelligent residence environments, intelligent abnormal behavior detection systems and human-robot interaction systems. Decision making behavior of intelligent agents is studied by many researchers and the result of these researches is proposed as various behavioral models. Lee et al. [1] categorized these models in 3 major approaches:

1. Economical approach
2. Psychological approach
3. Synthetic engineering-based approach

First, models in the economical approach have concrete foundation, mostly based on the assumption that decision makers are rational [2, 3]. However, one limitation is their inability to represent human cognitive natures. To overcome this limitation, models in the psychological approach (second category) have been proposed [4–6]. While they consider human cognitive natures explicitly, they mainly focus on the human behaviors under simplified and controlled laboratory environments. Decision Field Theory (DFT) is a famous model of this category. Finally, the synthetic engineering-based approaches employ a number of engineering methodologies and technologies to help reverse-engineer and represent human behaviors in complex and realistic environments [7–13]. The human decision-making models in this category consist of the proper engineering techniques employed for each sub-module. BDI, SOAR and ACT-R are widespread known models of this category. However, the complexity of such comprehensive models makes it difficult to validate them against the real human decisions [1].

In this paper a novel behavioral model is proposed which involves a decision making strategy and the agent's emotions as the most important factor in decision making process. Another novelty of this paper is that it utilizes a simple structure that any other affecting parameters such as agent's personality and memory can be augmented to in the future. The proposed model was tested on some agents in a goal reaching scenario.

2 Proposed Model

2.1 Main Idea

All living intelligent agents are consciously or unconsciously optimizing their lives. So every decision they make and every action they take is dedicated to this objective. Hence, we can conclude that decision making structure of every living intelligent agent includes a dynamic multi-objective goal function and an optimization structure. The goal function of every agent is specific and different from the others' and it is because of the differences in their objectives, personalities and other characteristics. But they are structurally similar and depend on the agent's emotions, feelings, morals, etc. The task of the optimization structure is to optimize the goal function in the manner of calculating the cost and benefit of every possible alternative at the decision making time and finally choose the best one which involves the most benefit and least cost. Meanwhile the moral, bodily and substantial characteristics and parameters like the agent's current emotional state interfere and affect

this optimization process so that the agent may make different decisions in the same situations.

2.2 Emotion Model

Emotions are a controversial topic and an important aspect of human intelligence and are shown to play a major role in decision making process of humans and some animals. Many scientists in the fields of psychology, philosophy and artificial intelligence proposed various models of emotion. Most of the proposed models focus on reactional behavior of the intelligent agent. However, through the history of emotion modeling, it has been shown that agent's other moral, substantial and bodily characteristics such as memory and expertise, personality, intelligence and physical situations play a major role in its decision making process too.

Ortony, Clore and Collins [14] proposed an emotion model, which is often referred to as the OCC model. There are also different emotion models presented from other researchers, such as Gomi [15], Kort [16], and Picard [17] and the FLAME model by Seif El-Nasr et al. [18]. Hidenori and Fukuda [19] proposed their emotion space. Wang et al. [20] also proposed another emotion space. Zhenlong and Xiaoxia [21] by combining the emotion space proposed by Hidenori and Fukuda [19] and the one proposed by Wang et al. [20] and based on the OCC model built their emotion space. Their emotion space includes four basic emotions Angry, Happy, Nervous and Relief. In this paper we apply their emotion space.

According to OCC model, emotions are caused by an agent's evaluation of an event. So, emotional state of an intelligent agent turns to a positive state if triggered by a positive stimulus and to a negative state if triggered by a negative one [22]. In the scenario of this paper the distance between the agent and its enemy (known as Enemy Distance) and the distance between the agent and its goal (known as Goal Distance) are stimuli. Goal Distance causes symmetrical emotions Happiness and Anger and the Enemy distance causes symmetrical emotions Nervousness and Relief. Fig. 1 illustrates our proposed circular emotion space of an intelligent agent.

2.3 Event Evaluation Fuzzy System (EEFS)

The task of Event Evaluation Fuzzy System (EEFS) is to map environmental stimuli into the agent's emotion space. This means EEFS determines which and how emotions are excited by events. This unit includes the following parts:

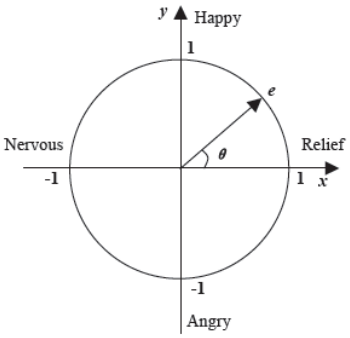

Fig. 1. Proposed emotion space of the intelligent agent

Input Variables. Enemy Distance (ED) with 9 membership functions (UC[1], VC, C, AC, M, AF, F, VF and UF) illustrated in Fig. 2 and Goal Distance (GD) with 9 membership functions (UC, VC, C, AC, M, AF, F, VF, and UF) illustrated in Fig. 3. This type of fuzzy partitioning of input space allows

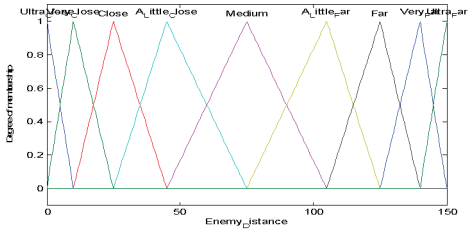

Fig. 2. Membership functions for input variable Enemy Distance

a slight nonlinear mapping of the input space to the output space. This is because of the nonlinear nature of emotion arousal in different situations.

Output Variables. Emotional Intensity trajectories x and y in Cartesian emotion space which both have 9 membership functions (UL, VL, L, AL, M,

[1] U=Ultra,V=Very,A=A little,C=Close,F=Far,M=Medium,H=High,L=Low

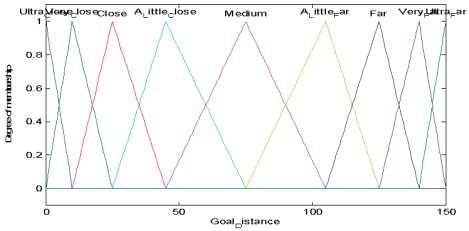

Fig. 3. Membership functions for input variable Goal Distance

AH, H, VH and UH) equally partitioning the output space ranging from -1 to 1 that one of them is illustrated in Fig. 4.

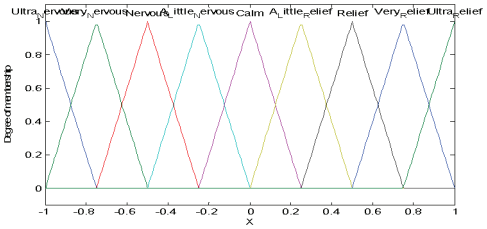

Fig. 4. Membership functions for output variables x and y

Fuzzy Rule Base. The rule base to manage the correlation between the inputs and the outputs of the EEFS is shown in Table 1.

Vector Representation of Emotions. The output of the EEFS are emotional intensity trajectories x and y in emotion space. So,

$$x = \{x | x \in \Re, -1 \leqslant x \leqslant 1\} \tag{1}$$

$$y = \{y | y \in \Re, -1 \leqslant y \leqslant 1\} \tag{2}$$

Here these variables form a square emotion space in a Cartesian coordination. For having a circle emotion space (like Fig. 1) we have to map these Cartesian

Table 1. Fuzzy rule base of emotion model

Rule No.	Goal Distance	Enemy Distance	y	x
1	UC	UF	UH	UH
2	VC	VF	VH	VH
3	C	F	H	H
4	AC	AF	AH	AH
5	M	M	M	M
6	AF	AC	AL	AL
7	F	C	L	L
8	VF	VC	VL	VL
9	UF	UC	UL	UL

coordination to a circular coordination.

$$x_c = x_s.\sqrt{1 - 0.5y_s^2} \tag{3}$$

$$y_c = y_s.\sqrt{1 - 0.5x_s^2} \tag{4}$$

Where x_s and y_s represent Cartesian coordination and $x_c$and y_c represent the new circular coordination representation. For simplicity we use x and y instead of x_c and y_c. On the other hand determining the type and the uniform intensity of the emotion is too hard having just these two numbers. So let us define Emotion Vector $\underline{e}$ as follows:

$$\underline{e} = [x, y] \tag{5}$$

In circular representation of emotions, emotion vector ($\underline{e}$) can also be represented by its Norm (ρ) and its Angle (θ).

$$\rho = \sqrt{x^2 + y^2} \tag{6}$$

$$\theta = \tan^{-1}(\frac{y}{x}) \tag{7}$$

Now we can simply define the intensity of emotions by the norm (ρ) and the type by the angle (θ) of emotion vector ($\underline{e}$). The correlation of the emotion angle, basic emotion, emotion intensity and final emotion is represented in Table 2.

For example emotional state $\underline{e} = (0.5, 30^\circ)$ is located in the first quadrant, its intensity is 0.5, its angle is 30°, so the corresponding emotion is Relief. When the agent's norm of the emotion vector is less than 0.2 we assume that its emotional state is Calm.

Table 2. Correlation of the emotion angle, basic emotion, emotion intensity and final emotion

Emotion Angle	Basic Emotion	Emotion Intensity	Final Emotion
$\frac{\pi}{4} \leqslant \theta < \frac{3\pi}{4}$	Happy	$0.8 \leqslant \rho \leqslant 1$	Very Happy
		$0.4 \leqslant \rho \leqslant 0.8$	Happy
		$0.2 \leqslant \rho \leqslant 0.4$	A Little Happy
$\frac{3\pi}{4} \leqslant \theta < \frac{5\pi}{4}$	Nervous	$0.8 \leqslant \rho \leqslant 1$	Very Nervous
		$0.4 \leqslant \rho \leqslant 0.8$	Nervous
		$0.2 \leqslant \rho \leqslant 0.4$	A Little Nervous
$\frac{5\pi}{4} \leqslant \theta < \frac{7\pi}{4}$	Angry	$0.8 \leqslant \rho \leqslant 1$	Very Angry
		$0.4 \leqslant \rho \leqslant 0.8$	Angry
		$0.2 \leqslant \rho \leqslant 0.4$	A Little Angry
$\frac{-\pi}{4} \leqslant \theta < \frac{\pi}{4}$	Relief	$0.8 \leqslant \rho \leqslant 1$	Very Relief
		$0.4 \leqslant \rho \leqslant 0.8$	Relief
		$0.2 \leqslant \rho \leqslant 0.4$	A Little Relief

3 Decision Making Strategy

Due to the structure of the field, the agent has 9 alternatives to choose between that consist of 8 alternatives (A,B,C,D,E,F,G,H) for moving in 8 directions and one alternative to stay in its current coordination (X). Fig. 5 illustrates these movement alternatives.

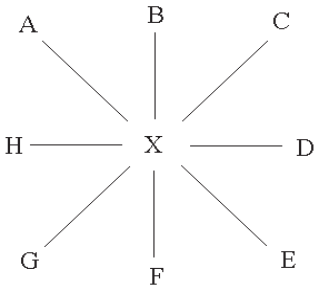

Fig. 5. Agent's possible movement alternatives

For building the Decision Making structure, first we need to define a Goal

Function to be maximized:

$$r^i = f_r^i - f_c^i \tag{8}$$

Here r is the Goal Function, f_r is the Reward Function and f_c is the Cost Function and index i represents the number to the corresponding alternative (X, A, B, C, D, E, F, G and H respectively). Reward Function determines the Reward of each alternative and the Cost function determines the cost of that alternative. The definition of Reward Function in our sample scenario is as follows:

$$f_r^i = e_c x^i + g_c y^i \tag{9}$$

Which e_c is the enemy prevention factor and g_c is the goal importance factor. This definition of reward function determines the agent approaches the goal and prevents the enemy. The factors e_c and g_c are dynamic control factors that depend on the current emotional state of the intelligent agent and will be discussed in the next section.

For a suitable definition of the cost function in our sample scenario, we need the definition of the energy consumed by each alternative:

$$f_c^i = e_k^i = \frac{1}{2} m (v^i)^2 \tag{10}$$

Where e_k is the kinetic energy, m the mass of the agent and v the velocity of movement. Here $m = 2$ and all kinds of friction is disregarded.

If the agent walks (makes one move per second) in orthogonal directions (B, D, F and H), its velocity is $v = 1$ units/sec so the energy consumed for this alternative is $e_k = 1$. Similarly if the agent walks (makes one move per second) in diagonal directions (A, C, E and G), its velocity is $v = \sqrt{2}$ units/sec so the energy consumed for this alternative is $e_k = 2$. Staying in the current coordination (X) does not consume energy. On the other hand running (making two moves per second) in every direction doubles the velocity, leading into 4 times energy consumption.

Now we are ready to recast and complete the goal function defined by (8), (9) and (10):

$$r^i = e_c x^i + g_c y^i - \alpha e_k^i \tag{11}$$

α is a dynamic factor as energy saving importance factor which depends on the personality and the physical situation of the agent and will be discussed in the next section. So the decision making strategy would be as follows:

$$i^* = Arg(\underset{i}{Max}\, r^i = e_c x^i + g_c y^i - \alpha e_k^i) \tag{12}$$

4 The Role of the Agent's Emotions in Its Decision Making Process

The decision making strategy proposed by (12) leads to a deterministic and optimal agent behavior in our sample scenario. But living intelligent agents do not necessarily make optimal decisions. In living intelligent agents no decisions are made isolated and without any interferences and moderations by its emotions. The agent's emotions play an important role in its decision making process. For instance it is obvious that the decisions made by a nervous person are different from the decisions made by that person when he/she is in a relief emotional state. This means the behavior of intelligent agents are to some extent stochastic rather than being completely optimal and deterministic. Therefore, we have to add the influence of emotions to our decision making strategy defined by (12). This can be achieved by changing dynamic factors e_c and g_c and so, it will lead to more believable, intelligent and natural agents.

The factor e_c is enemy prevention factor. Intensity of nervousness increases this factor and so the agent's tendency to escape from enemy. Meanwhile, g_c or the goal achievement importance decrease, so leads to the agent's less tendency to reach to its goal. So, in nervous emotional state:

$$\begin{cases} e_c = \rho \\ g_c = 1 - \rho \end{cases} \tag{13}$$

ρ can be obtained by (6).
On the other hand, the reverse procedure happens when the agent approaches near its goal. So

$$\begin{cases} e_c = 1 - \rho \\ g_c = \rho \end{cases} \tag{14}$$

In other emotional states:

$$\begin{cases} e_c = 1 \\ g_c = 1 \end{cases} \tag{15}$$

In addition to the above mentioned influences, the emotional state of the intelligent agent – in particular when the agent is under a high amount of stress – affects its decision making process in another way. Stress causes the agent to decide incorrectly. The strategy defined by (12) always returns the optimal alternative (i^*). The optimal solution can be obtained by the following equation:

$$i^* = Arg(\underset{i}{Max}(r^i)) \tag{16}$$

Now we have to show the effect of stress in its decision making process. To enclose the influence of stress we can use Quasi-Boltzmann's probability equation as follows:

$$p^{i^*} = \frac{1}{1 + e^{(-\frac{1}{|x^0|})}} \quad , x \leqslant 0 \tag{17}$$

Here p^{i^*} is the probability of choosing the optimal solution and x^0 is the emotion intensity's x-axis trajectory of current emotional state. Regarding (16) if the agent's emotional state is not nervous ($x^0 \geqslant 0$) the probability of choosing the optimal solution is 100%, and if its emotional state is very nervous ($x^0 = -1$), the probability is 73.11%. So in this situation the agent may choose a wrong alternative and get caught by the chasing enemy.

By adding emotions, the final model of the agent's decision making strategy is constructed. The block diagram of the agent's decision making structure is illustrated in Fig. 6.

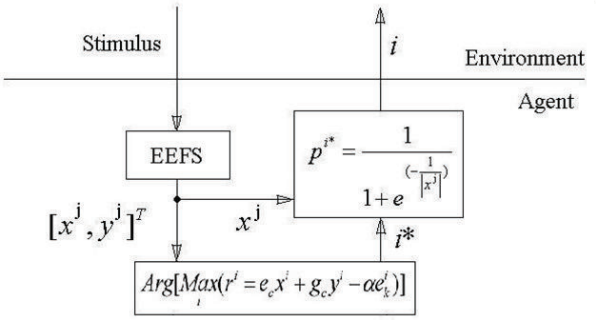

Fig. 6. Block diagram of the agent's decision making structure

5 Simulation

As mentioned before, the sample scenario of this paper includes an agent and its goal and enemy. The aim of the agent is to reach to its goal with minimal energy consumption while preventing to be hunt by its enemy. The field is square with 100 by 100 allowed points. Both agent and enemy are just allowed to move orthogonally and diagonally.

Some examples of simulated behavior of some similar agents are shown in Fig. 7-10. The Green Star represents the location of the Goal; the Red Star represents the starting point of the enemy; the Magenta point and square represent the starting point of the agent; Yellow points represent the enemy path when the agent is not in its eyesight (Enemy Distance $>$ 30m); Magenta points represent the agent path when it is feeling "Very Nervous" (Enemy Distance $<$ 18.5m) and is escaping from the enemy and also represent the enemy path while chasing the agent; Cyan points represent the agent path when its emotional state is anything other than the state "Very Nervous"; Red points represent the agent path when it is tired ($e_k \leqslant \lambda = 25\%$) and finally Blue points represent the wrong decisions made by the agent when it feels "Nervous". For maximizing the believability of the model, we defined energy consumption for the enemy so after a certain chasing duration, the enemy feels tired and will not start chasing the agent unless its energy is higher than a certain threshold. Also as can be seen, because the enemy has a hunter personality, its eyesight power to start chasing (30m), is greater than the eyesight of the agent when it feels "Very Nervous" and starts escaping from the enemy (18.5m).

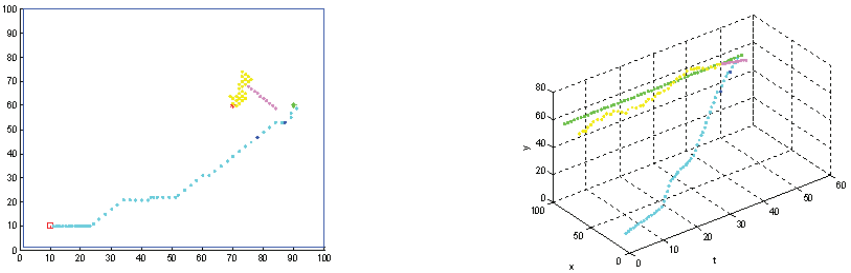

Fig. 7. Behavior of agent #1 in 2D view (left) and 3D (versus time) view (right)

6 Conclusion and Future Work

In this paper a novel model of behavior for intelligent agents was introduced and its validity was examined on four similar agents in a goal-approaching scenario with minimal energy consumption and maximum enemy prevention.

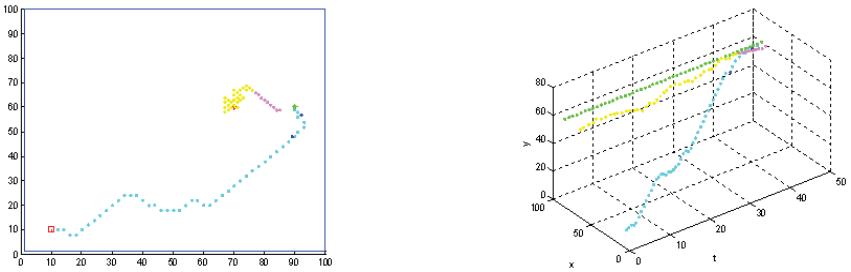

Fig. 8. Behavior of agent #2 in 2D view (left) and 3D (versus time) view (right)

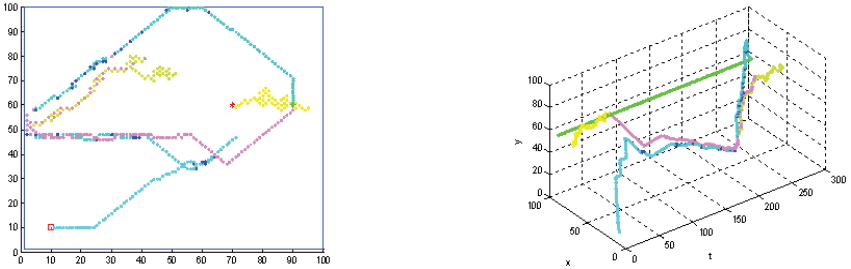

Fig. 9. Behavior of agent #3 in 2D view (left) and 3D (versus time) view (right)

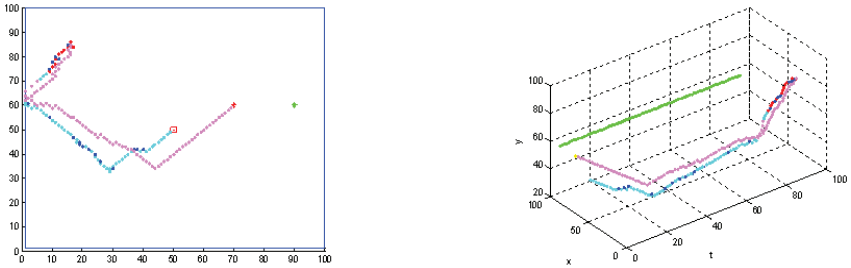

Fig. 10. Behavior of agent #4 in 2D view (left) and 3D (versus time) view (right)

Regarding Fig. 7-10, the agent's behavior in this scenario is intelligent, natural and believable. Also the effect of the agent's emotion on its behavior was obvious. For example, as can be seen in Fig. 10, the agent #4 at coordination (17, 86) faced lack of energy and also made a wrong decision because of its "Nervous" emotional state and so got hunt by its enemy.

The decision making strategy of this paper is proposed based on four basic emotions, but any other emotions can be augmented to the model (Eq.12) easily. Augmenting any other bodily, substantial and moral characteristics to the model can be easily achieved too.

Still much amount of research and development is required in order to obtain a complete and comprehensive model. Applying more complex scenarios, simulation in a multi-agent environment and combining this model with other intelligent methods such as Artificial Neural Networks, Reinforcement Learning and Evolutionary Algorithms could be the horizons for future works.

References

1. Lee S., Son Y.J.: Integrated Human Decision Making Model Under Belief-Desire-Intention Framework For Crowd Simulation. In: Mason, S. J. , Hill, R.R., Mönch, L., Rose, O., Jefferson, T., Fowler, J. W. (eds.) Proceedings of the 2008 Winter Simulation Conference (2008)
2. Opaluch J. J., Segerson K.: Rational Roots of Irrational Behavior: New Theories of Economic Decision-Making. Northeastern Journal of Agricultural and Resource Economics 18(2), 81–95 (1989)
3. Gibson F. P., Fichman M., Plaut D. C.: Learning in Dynamic Decision Tasks: Computational Model and Empirical Evidence. Organizational Behavior and Human Decision Processes 71, 1–35 (1997)
4. Einhorn, H. J.: The Use of Nonlinear, Noncompensatory Models in Decision Making. Psychological Bulletin 73, 221–230 (1970)
5. Payne J. W.: Contingent Decision Behavior. Psychological Bulletin 92, 382–402 (1982)
6. Busemeyer J. R., Townsend J. T.: Decision Field Theory: A Dynamic-Cognitive Approach to Decision Making in an Uncertain Environment. Psychological Review 100(3), 432–459 (1993)
7. Laird J. E., Newell A., Rosenbloom P. S.: SOAR: An Architecture for General Intelligence. Artificial Intelligence 33, 1–64 (1987)
8. Newell, A.: Unified Theories of Cognition. Harvard University Press, Cambridge, Massachusetts(1990)
9. Rao A. S., Georgeff M. P.: Decision procedures for BDI logics. Journal of logic and computation 8, 293–342 (1998)
10. Konar, A., Chakraborty U. K.: Reasoning and unsupervised learning in a fuzzy cognitive map. Information Sciences 170, 419–441(2005)

11. Zhao, X., Son Y.: BDI-based Human Decision- Making Model in Automated Manufacturing Systems. International Journal of Modeling and Simulation (2007)
12. Rothrock L., Yin J.: Integrating Compensatory and Noncompensatory Decision Making Strategies in Dynamic Task Environments. In: Kugler, T., Smith, C., Connolly, T., Son, Y. (eds) Decision Modeling and Behavior in Uncertain and Complex Environments, pp. 123–138, Springer(2008)
13. Lee S., Son Y., Jin J.: Decision Field Theory Extensions for Behavior Modeling in Dynamic Environment using Bayesian Belief Network. Information Sciences 178(10), 2297–2314 (2008)
14. Ortony A., Clore G., Collins A.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge, UK (1988)
15. Gomi, T., Vardalas, J., Koh-Ichi I.: Elements of Artificial Emotion. In: Robot and Human Communication, pp. 265–268 (1995)
16. Kort, B., Reilly, R., Picard, R.W.: An affective model of interplay between emotions and learning. In: Proceedings of IEEE International Conference on Advanced Learning Technologies, pp.43–46 (2001)
17. Picard R.W., Vyzas E., Healey J.: Toward Machine Emotional Intelligence-Analysis of Affective Physiological State. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 1175–1191 (2001)
18. Seif El-Nasr M., Yen J., Ioerger T.R.: FLAME-Fuzzy logic adaptive model of emotion. International Journal of Autonomous Agents and Multi-Agent Systems (2000)
19. Hidenori I., Fukuda T.: Individuality of Agent with Emotional Algorithm. In: Proceedings of IEEE 2001 International Conference on Intelligent Robots and Systems, pp.1195–1200 (2001)
20. Wang Z., Qiao X., Wang C., Yu J., Xie L.: Research on Emotion Modeling Based on Custom Space and OCC Model. Computer Engineering 33(4), 189–192 (2007)
21. Zhenlong L., Xiaoxia W.: Emotion modeling of the driver based on fuzzy logic. In: 12th International IEEE Conference on intelligent Transportation Systems. St. Louis, MO, USA (2009)
22. Chakraborty A., Konar A., Chakraborty U. K., Chatterjee A.: Emotion Recognition from Facial Expressions and its Control using Fuzzy Logic. IEEE Transactions on Systems, Man, and Cybernetics (2009)

A Visit on the Uncanny Hill

Petr Švarný

Charles University, Prague, Czech Republic
svarnyp@logici.cz

Abstract. The article introduces shortly the Uncanny valley hypothesis and sums up some of the research done in the field connected to it. It explores the possible new options in research or robot design which could help to subdue this uncanny obstacle on the way to a pleasant human-robot interaction. It also mentions the possible risk of an uncanny valley phenomenon the other way around, from the view of artificial intelligence (AI).

Keywords: human-robot interaction, uncanny valley

1 Introduction

This paper explores the so called Uncanny valley hypothesis in the light of the use of humanities and art in human-robot interaction. As all sorts of AI systems take a bigger part in our day to day lives, we more often face the question how to make human-robot interactions pleasant and natural-seeming. This problem was studied already in 1970 by M. Mori [1, 2], who introduced the hypothesis how people react to human-like entities. We will describe this hypothesis briefly and show some results concerning its verification. Thereafter we focus on possibilities how the hypothesis of an uncanny valley could be treated with inspiration coming from art. We suggest that the valley should be approached from the side of the AI also.

2 The Valley Ahead

The Uncanny valley hypothesis claims that the familiarity, affinity, or comfort of our contact with an entity that is similar in some respects to humans is not a simple linear function. Although it is true that the more human-like an entity is, the more we are comfortable while interacting with it, Mori supposed that there is a sudden drop in comfort as we reach a certain point of realism

and it does not cease unless we face a real human entity. According to this hypothesis, a human test subject should feel little affinity towards robots that are not similar to humans (see industrial robots). The subject should have some level of affinity to humanoid robots but should have an eerie sensation when confronted with an actroid[1]. It was already in the original article that the difference between motionless and moving entities was explored. Mori mentioned the different feeling we have when facing a simple prosthetic arm that is still and when we observe a myoelectric hand[2].

The topic caught more attention today than at the time when the article was published. Generally the hypothesis finds support in today's research. For example, we can see the attempts to broaden the studied aspects in [3]. However, there is present also an opposite view. We can take [4] as an example of an article that tries to eliminate the valley. Medical investigations also could be taken into account as prosopagnosia or the way how we react to first time exposure to unfamiliar faces might play an important role in the subject.[3]

3 Valley Hiking in the Modern World

One of the main questions to answer before we try to venture into the valley is, if it is necessary to climb up the hill to realism and affinity. A good artistic example of this could be Johnny 5 – he has rudimental options how to express emotions, he is not human-like but has some basic human characteristics, and he reacts similarly as a human being would. He represents a robot that is comfortable to interact with, although he does not have human-like features.

However, Hanson et al. present the following reason why it is worth trying to achieve realistic human robots:

> ... realistically depicted humanlike robotics will serve as an unparalleled tool for investigating human social perception and cognition. In our experiments, our robots have demonstrated clearly that realistic robots can be appealing. We conclude that rendering the social human in all possible detail can help us to better understand social intelligence, both scientifically and artistically. [4] (p. 31)

[1] An android that is visually very human-like.

[2] Basically a moving prosthetic arm. The mentioned example is directed by electric signals received from human skin surface.

[3] See for example [5] showing that basic observation of facial behaviour is deep-rooted and it is present already at a very young age. The great speed with which people react to facial stimuli is shown in the study [6].

This quote mentions social perception and cognition. Therefore, we can point out one of the possible problems connected to the studies of the uncanny valley – they do not use commitment and longer term cooperation. These are present in many human interactions and often play an important role in the formation of our social life. Any feelings of eeriness and discomfort connected to human-like robots could possibly vanish after a few days of interaction and be replaced with genuine affection.

However, we might not need realistically humanoid robots in order to have a comfortable human-robot interaction. As the first idea coming from art, we mention McCloud's observation from the art of drawing comics. He claims [7] (p. 31) that simple shapes allow the reader for more immersion as they allow for more universality. Any character that is depicted in a realistic manner is understood by the reader automatically as something different, something exterior to which he cannot relate that easily. This takes into account also the human tendency to recognize faces in many simple shapes (for example due to pareidolia) and allows us to construct robots with simple forms of facial expressions. Nevertheless, we need to pay attention to the fact that the immersion present in comics is due to some other factors also: we are often the witnesses of the character's thoughts, the character is expressing emotions, and she is reacting to the situations she faces in an unsurprising way. This would suggest that a successful comics based interaction is given by a robot that has a simple facial interface and reacts in a way we would expect it to react.

We can drop the option to share inner thought processes for two reasons. First, it is a common and quite accepted response in a conversation between people to answer: "I don't know", when one is asked about a difficult thought process. Second, if the robot achieves the other two mentioned points, it will be attributed a mind by his human colleagues.

We cannot leave the other two demands aside. Being confronted with humanoid robots that do not react in an expected way can be similar to facing a human that reacts abnormally. It leads to a reaction of fear and panic because the theory of the mind of the encountered person fails to predict or explain his actions. The fact that unexpected behaviour is alien to us already from early age is shown for example in [5]. Infants react strongly if their communication counterpart does not follow the usual pattern of behaviour and suddenly stops reacting to stimuli.

For the second demand, if we would not request a simple facial interface, we would return to the original idea of trying to make human-like robots instead of making only robots that are pleasant to interact with or we would remain with machine-like robots. At this point it is our main concern to ameliorate

the interaction between humans and robots at the lowest cost possible. If we focus on facial realism, we might end up with a machine that is great at expressing emotions but is too complex for a daily use in our lives. On the other hand, if we omit facial features altogether we fail to facilitate the human-machine interaction. For this reason we want to stay with a design as simple as possible.

In many respects the fact that human communication is nowadays often also dependent on a computer interface facilitates our attempts to befriend humans with robots. Many people grow up expressing their emotions in emoticons and text messages and receiving emotional responses in a similar way. A recent movie named Moon has shown a robot called Gerty that communicated with an emotionally neutral voice but his statements were accompanied with an emoticon on his main screen showing his mood. It was thanks to this small screen that communication with Gerty seemed much more pleasant than communication with HAL9000 from the movie 2001: A Space Odyssey.

Many other interactions do not even need any visual interface to work properly. Already the old psychoanalysis program called Eliza has proven somewhat effective in fooling people into believing she had some mind or intelligence, although she had none [8]. A modern counterpart of Eliza is Apple's Siri, an intelligent personal assistant that responds to voice commands and reacts only in voice or by giving the demanded output behaviour (for example, sending an email). Obviously such applications do not fall into the uncanny valley, but they show how minute the trouble with the valley can be. Emotional modulation of the AI's voice could be enough to give people (already used to talking over phones) enough feedback to make the interaction close to a human-human exchange. The crucial point is the difference in importance people ascribe to visual and auditory stimuli. In order for the conversation to meet our two demands, the robot could even have a static chassis and demonstrate all its reactions by his audio systems. This view also leads to the important question of application. What would be the use of a human-like realistic robot?

As the subtitle of the conference is "artificial dreams", the reference to P. K. Dick's "Do androids dream of electronic sheep?" comes into mind. The human-like androids in that world are used for mining and similar labour. Such use seems simply unrealistic as it would probably be more cost effective to have specialized machines for these purposes. The scenario of personal assistants is a more realistic and probable one. Following in the footsteps of Siri they could take the form of an audio responding humanoid with supressed or simplistic and non-changeable facial features. We return here again to the question if the valley needs to be crossed. Employing a realistic humanoid assistant would

only lead to affinity towards this assistant and possible impairment on the effectiveness of its use (for example one would want his assistant to take some rest or go for vacation). On the other hand, a well-designed assistant – let us say still on the hill before the steep drop into the valley – could already make its human user comfortable enough but prevent him from ascribing too many human characteristics to the assistant. This could be achieved by maintaining an illusion of correct emotional response and simplistic representation.

4 Foreign Visitors to the Valley

We focused the whole time on the human-robot interaction. If we imagine, however, a robot already capable of genuine emotional response, we can ask also about the robot-human interaction. If there is a human-robot uncanny valley, would there be also one for the artificial participants in the conversation? How would their emotions react to other robots, perceived by humans as uncanny? Obviously it is a question closely tied to the mechanisms that would be incorporated into these robots and thus for now unanswerable.

However, it might already be the time to start evaluating whether we shouldn't prepare artificial/AI/robot equivalents of some humanities. Especially psychology could be transformed into a tool to work with AI from a top-down perspective. This might need to be as specialized as its human counterpart and couldn't be simply presented as some interdisciplinary effort between psychology and AI. A more "biological" approach to robots and AI could also help to classify any eeriness or bizarre behaviour as AI counterparts of human abnormal states without getting lost in too complex bottom-up descriptions and at the same time it would allow the treatment of AI in a similar manner as humans or animals are treated. A good example of a topic from psychology that could be useful for our cause is the Asperger syndrome. A person suffering from this disorder might often make other people uncomfortable and thus slip into the uncanny valley.

The ultimate use of many of the here mentioned ideas – even the use of non-human like assistants or psychological classifications – is closely tied to the ethics of AI. Do we want to ascribe the same status to beings evolved from human research and effort as to those that evolved from the chaos of the universe?

5 Conclusion

We have introduced the idea of the uncanny valley from M. Mori that robots that are human-like might make people feel eerie because of their imperfect

similarity to humans. We suggested that the valley does not have to be taken as an obstacle with regards to the design and goals of many AIs and robots even if they would be interacting with people on a daily basis. Some questions still need to be answered before the valley could be left for good. What stimuli are more relevant in human-human interaction? Aren't contemporary humans already used to computerized interactions? If so, is it enough to overcome the valley and make interactions with robots comfortable? Shouldn't a holistic approach, as AI-psychology, be introduced into AI to deal with similar problems?

References

1. Mori, M.: Bukimi no tani. Energy 7(4), 33–35 (1970)
2. Mori, M., MacDorman, K.F. da Kageki, N.: The uncanny valley [from the field]. Robotics & Automation Magazine 19(2), 98–100 (2012)
3. Ho, C., MacDorman, K.: Revisiting the uncanny valley theory: Developing and validating an alternative to the godspeed indices. Computers in Human Behavior 26(6), 1508–1518 (2010)
4. Hanson, D., Olney, A., Prilliman, S., Mathews, E., Zielke, M., Hammons, D., Fernandez, R., Stephanou, H.: Upending the uncanny valley. In: Proceedings of the National Conference on Artificial Intelligence. Volume 40(4). AAAI Press, MIT Press (2005)
5. Cleveland, A., Kobiella, A., Striano, T.: Intention or expression? four-month-olds reactions to a sudden still-face. Infant Behavior and Development 29(3), 299–307 (2006)
6. Hadjikhani, N., Kveraga, K., Naik, P., Ahlfors, S.: Early (n170) activation of face-specific cortex by face-like objects. Neuroreport 20(4) (2009)
7. McCloud, S.: Understanding comics: The invisible art. Harper Paperbacks (1993)
8. Weizenbaum, J.: Eliza—a computer program for the study of natural language communication between man and machine. Communications of the ACM 9(1), 36–45 (1966)

Why Are We Afraid of Robots? The Role of Projection in the Popular Conception of Robots

Michael Szollosy

The University of Sheffield, Sheffield, UK
M.Szollosy@sheffield.ac.uk

Abstract. The popular conception of robots in fiction, film and the media, as humanoid monsters seeking the destruction of the human race, says little about the future of robotics, but a great deal about contemporary society's anxieties. Through an examination of the psychoanalytic conception of *projection*, this essay will examine how robots, cyborgs, androids and AI are constructed in the popular imagination, particularly, how robots are feared because they provide unsuitable containers for human projection (unconscious communication) and how at least part of what we fear in robots is our own human rationality.

Keywords: robots, cyborgs, AI, psychoanalysis, projection, uncanny valley

I come from a background teaching cultural studies and psychoanalysis. When I started working with the Sheffield Centre for Robotics, I was charged with this, rather straight-forward, question: Why are we afraid of robots? If we look at the cultural evidence, from literature, film and video games, and in the popular media, it seems that robots have entered the popular imagination as monsters on scale comparable to the vampire (and also, it should be noted, with a similar level of ambivalence[1]). However, perhaps predictably, there is no single, simple answer for a phenomenon so ubiquitous, no single theory that will explain why we are presented again and again with humanoid machines that want to attack, subvert or enslave the human race. What is

[1] We are not, of course, afraid of all robots. There are some robots with which we have a very different set of relations. These are largely, I still maintain, based on projections, though a very different set of projections than those which I am about to describe in this paper.

evident is that, as with most of humanity's monsters, the way we perceive robots says much more about our own anxieties now than any real present or future developments in robotics. It is my hope that a thorough analysis of how robots are portrayed in popular imagination can not only help us better understand these underlying anxieties and fears but also inform those designing the robots of the future as to how their inventions might be met by the public.

To the question, why are we afraid of robots, I want to propose at least two, intricately related ideas here:

1. We are afraid of the robot because of the existential threat it represents to our humanity. But by this I must emphasise that I *do not* mean that we genuinely fear robots will arise with their familiar arsenal (deception, fantasy machines, laser blasters) and wipe humanity off the earth, as it is so often imagined. Rather, this threat lies in our own fantasies and conceptions of ourselves, notions that I best understand and can explain through the notion of *projections* – complex psychological processes of relating described in psychoanalytic clinical and cultural theory. Robots, and humanoid robots in particular, are regarded (not without good reason) as empty, unyielding *containers* that cannot give or take or function in the normal course of human projections. They are incapable of receiving projections, which in more general language means that they are incapable of *empathy*, but understood through the idea of projections we can grasp the consequences of this in much greater detail. The humanoid robot, therefore, is instead transformed into a menacing, persecuting figure that becomes a container for all of our own negative emotions – the hate and violence of the robot is our own hate and violence that we imagine is out there, characteristic of these monsters instead of ourselves.

2. From this, it is apparent that our fear of robots is at least in part a fear of our own *rationality*, dead, mechanical and calculating. Both the robot and reason are humanity's own creations, inventions that we fear are becoming autonomous monsters more powerful than their creator. Somewhere, too, in that *simulacra* of humanity – this robot that we have created in our image, that looks like us and comes to represent us to ourselves – we are afraid of losing the very qualities that we deem define us as human. We fear becoming that empty shell of cold, mechanical, unfeeling rationalism. Like so many of our monsters, from Frankenstein to andys [1] to the Terminator [2], the Borg [3] and even Wallace's wrong trousers [4], we

fear what we have created, and we fear that the process of construction – that science itself – will render us less human.

These ideas, I believe, also provide a more detailed account for the phenomenon of *the uncanny valley*, an idea which, after all, has at least a certain root in Freud's early psychoanalytic thinking, and evidence for some of this way of regarding robots and our technological future may be found in the debate between the 'transhumanists' and their self-styled nemeses, the 'bioconservatives', and I hope to make some remarks upon this at my conclusion.

Projection is an idea with its roots in Freudian psychoanalysis, but has been considerably enriched by Freud's disciples and contemporary psychoanalytic clinical and cultural theory. The concept of projection tries to describe *object relations*, that is, the way that people relate to things – usually other people, but also other material and non-material objects in their world. Ideas of projection, and the related notion of projective identification, are used in cultural studies to provide compelling explanations for phenomenon as diverse as Nazism and teenage crushes, racism and sports spectatorship.

In projection, it is believed that in psychological fantasy we split off parts of ourselves and 'project' them into something else – a person, an object, or even a symbol or an idea – which can then be regarded as a sort of *container* for these projections. Sometimes, good parts of the self are projected into containers, for safe keeping, for example, or in the case of projective identification, one may project a good part of the self into a container so that it can identify with that part in another. This idea of projective identification is the basis for *empathy*, but also provides a compelling explanation for cultural phenomena such as nationalism, for example, wherein individual people project their own positive qualities (say, resilience) into a symbol, or an idea, or a leader. When a number of people all identify with positive qualities projected into the container, it provides a collective cohesion, a group identity.

On the other hand, sometimes negative parts of the self can be projected into a container (and in practice it is usually a combination of good and bad parts that are projected). Bad parts of the self – violent fantasies, hatred, for example – can be projected away from the self, in order that the self can be thought of as pure and all good. When such projections find a home in another, that other then becomes the source of that badness, and becomes a persecuting figure as the hatred and violence that is projected out is now imagined returning in the form of the other. The most obvious examples of such projections are instances of scapegoating, such as commonly seen with racism (and here we see another all-too-familiar component of nationalism): It is not we who are violent, it is *them*. They hate us and are out to get us.

As with the scapegoat, there is a belief that the container of the bad parts of the self must be destroyed before it can return and destroy us. This is a root of paranoia. The belief that we are being persecuted is our own fantasy.

Though Freud introduced the notions of projections, more contemporary psychoanalytic thinking has elevated this idea to greater, or even of the utmost, importance. Projections and projective identifications are, for many, at the very centre of human communications and human experience, driven by what is described as an *epistemophilic impulse*, a desire to know [5]. Projections are a way of managing the anxiety aroused by the unknown, a fear of the other, which is particularly important in our investigations into robots. It is through such projections that we come to know and understand the world, through reality testing and an emotional engagement with the world. Into an unknown, uncertain space, we fantasise all sorts of things in order to defend ourselves against the greater fear of uncertainty and emptiness. The baby, psychoanalysts claim, will look at his mother as a mysterious, unknown other. In happy, or at least normal, times the baby might imagine in his mother a healthy mix of good and bad objects and motives. However, at times – and this is true even in normal development – the baby projects his own bad objects, his anger and frustrations, into the mother. Those bad parts are now imagined to originate and reside in his mother. The baby will imagine therefore his mother as the source of all present and future threats to its being. The baby, psychoanalysts predicted, would regard these threats contained in the mother in concrete terms, as poo and other babies lurking within the mummy, waiting to be expelled or unleashed to destroy the baby and his world. By imagining such things and placing them inside the unknown space, the baby acquires a sort of mastery over the unknown, and over his mother – he now knows what is there, because he has put it there. This has the consequence, however, of making this other space the source of badness, a place of evil. It becomes something that returns to persecute, to attack – but, again, this is only the baby's own imagination reflected back onto himself; he imagines his own violence, now out there, will come back to get him.

For an example of this as a cultural phenomenon, we need look no further than the fantasies of imperialism throughout history. European explorers in the nineteenth century, faced with the dark, unknown hearts of continents, used their imaginations to populate them with all sorts of savages, cannibals that always acted violently and without a trace of reason, while the 'civilised' Europeans themselves committed genocide and plundered resources. These imagined savages were nothing more than the darkest, most violent impulses of the imperialists projected out onto the external others, demonised to justify violent oppression, war and mass murder, and by keeping these bad parts of

themselves away and projecting them into another it simultaneously allowed the empire to believe its intentions noble, maintaining the ideal fantasy of empire as civil and good. (Unfortunately, we still see such processes at work in some historical accounts of European imperialism, and also in contemporary neo-imperialist practices.)

We see the same processes at work in the construction of our monsters throughout the ages, and now we see the same processes in popular representations of robots. The Terminator, for example, or *Star Trek*'s Borg are, among other things, projections of our own, very human, violent fantasies projected onto an other, an other which then becomes a relentless, supremely destructive persecuting object. In *Do Androids Dream of Electric Sheep?*, Phillip K. Dick's novel that is the basis of Ridley Scott's *Blade Runner*, the main character, Rick Deckard, provides us with a terrific example of how such projections operate. The bounty hunter, the epitome of the loner, Deckard nevertheless believes that it is the humanoid robot – the 'andy' – that is 'a solitary predator'. The narrator tells us, 'Rick liked to think of them that way; it made his job palatable' [1], which demonstrates how projections can function not just through an individual but as the founding myth of an entire profession, e.g. the bounty hunter, the police, or even an entire culture. Referring to the dominant spiritual and moral system of earth in this future world, Mercerism, the narrator explains how projections function as a defence, to maintain an idealised humanity while justifying murder and violence:

> In retiring – i.e. killing – an andy he did not violate the rule of life laid down by Mercer. [...] A Mercerite sensed evil without understanding it. Put another way, a Mercerite was free to locate the nebulous presence of The Killers wherever he saw fit. For Rick Deckard an escaped humanoid robot, which had killed its master, which had been equipped with an intelligence greater than that of many human beings, which had no regard for animals, which possessed no ability to feel empathic joy for another life form's success or grief at its defeat – that, for him, epitomized The Killers. [1]

Thus projected, the violence is not Deckard's own – it is the andys, The Killers, who are violent; it is their impulses that must be contained. These projections allow Deckard to reason that his own violence, the 'retiring', or murder, of the andys, is the only rational response to such seemingly external violence.

For many psychoanalysts, projection and projective identification are simultaneously the basis of all normal human development and inter-subjective

communications *and* for psychopathology and virulent cultural practices (fascism, imperialism, racism, etc.). For some, as well, the idea of projection is part of normal development and 'reality testing' in a way akin to the idea of 'negative feedback' in cybernetics [6]. The difference between 'normal' and 'abnormal' or 'pathological' in this case is a matter of degrees – uncomfortable distinctions, yes, but ones that need to be made nevertheless. As Robert Young says, 'What is crazy and murderous and what is essential to all experience and human relations are the same. *The same*' [6].

Projections provide a defence, as we have seen above, against unwanted parts of the self. Such fantasies are key to our understanding of self, and to maintaining a coherent sense of being, a psycho-somatic integrity. It is in these contexts that robots can represent an *existential* threat. Psychoanalysts believe that excessive splitting and projections can leave one feeling fragmented, in pieces. Projections can also be 'misplaced', that is, projected into an unsuitable container, one that is incapable of returning the projections in a useful way, offering feedback and confirmation of the fantasy. Such unsuitable containers can cause a feeling of being depleted and weakened, which can lead to a sense of futility and lacking feeling. Such sensations are referred to as *depersonalisation*, a feeling of not being real, which psychoanalysts sometimes describe as being akin to feeling like an automaton, an empty object in a world of empty objects [7], [8].

Robots are often portrayed in film and literature as being at their most dangerous when they are indistinguishable from humans – again, recall *The Terminator* films, the remake of *Battlestar Galactica* or Dick's *Do Androids Dream?*, where the questions of flesh or machine are paramount. Deckard, along with the rest of the human population, longs to keep real animals, not mechanical imitations; it is feared that andys live hidden in plain view amongst the human population, and it is Deckard's job is to distinguish between them. The fear that we cannot tell the difference between man and machine is an existential fear, not just in that that we cannot identify, literally, what it is that is 'human' and what is a copy, but that we are unsure who to trust with our projections. An unsuitable container can have dire consequences for the integrity and conception of the self. This is demonstrated in *Do Androids Dream?*: Deckard very explicitly explains that it is this inability to receive his projections that, at least in part, is responsible for his hatred of andys:

> He thought, too, about his need for a real animal; within him an actual hatred once more manifested itself toward his electric sheep, which he had to tend, had to care about, as if it lived. The tyranny

> of an object, he thought. It doesn't know I exist. Like the androids, it had no ability to appreciate the existence of another. [1]

Furthermore, we can see here the existential threat posed by this mere 'object' – it doesn't know he exists. The electric sheep, like the android, is incapable of confirming his existence by relating to him through projections. Projections must be seen to have consequences; they must be processed, returned, or spurned in some way. The android, however, like the 'dead mother' of psychoanalytic literature [9], is incapable of returning projections. Projections made into the android or the electric sheep are lost, devoured by the cold, unresponsive machine.

The theory of *the uncanny valley* has long maintained that it is the robots that look *most human* that are regarded to be the most dangerous. But why? The idea of projection provides us with another answer (not necessarily to discount any others): because it is when robots appear human that we are tempted to engage with them as humans and not as machines. When it approximates a human we are able to project those parts of ourselves that make us vulnerable to depersonalisation and disintegration; this is when the limitations of the machine threaten our own self, the fabric of our being. This returns us to Freud's initial notion of the Uncanny: what threatens us is the *unthought known*, the reflection of self that we cannot accept as the self, that we dare not acknowledge.

Furthermore, and this I shall return to in my second point, humanoid robots remind us how close we are to inhumanity ourselves – not that, as some would hold, they remind us of our own mortality, but that they show us what we might become: depersonalised, depleted of affect, empty of those good parts of the self that enable us to empathise and engage with the world beyond reason.

It is a question of *use*. We are happy to use robots to perform for us, as entertainment, or to work for us, as slaves. We even might use robots at times as a substitute when we wish precisely *not* to engage with the world, as a defence from the vicissitudes of emotional engagement. But when we are invited to use the robot as a container for those parts of ourselves that are more vital to our very notion of self, we balk, we recoil. We recognise it as an unsuitable container for the good parts of ourselves. The robot instead becomes a container for our negative emotions, those parts of ourselves that we want to dissociate from ourselves. But we fail to see that fear and anxiety and violence as our own and imagine instead that it originates from the robot itself. Thus, the robot becomes our creation not only in its physical construction but also in its 'programming', if you will – not just the instructions that we give

it how to behave, but in our imagination. Our own darkest impulses and fear become displaced onto the machine. We imagine that it wants to destroy us. It becomes a persecuting object. It is the machine that is driven by insecurity to destroy what it thinks threatens it. It is the machine that seeks vengeance. It is the machine that is driven by lust for conquest and empire.

Does the machine feel any of this? Of course not. But the robot/android has become another of humanity's great monsters – like the vampire, or the zombie, or so many other, more culturally specific beasts (which are so often the victims of scapegoating). We construct these monsters in our minds. They become containers for all of those feelings – our feelings, projections on to this external other so that we can imagine these impulses are not our own but theirs, something that belongs *out there*, *to them*, not our own, not lurking within us.

And this leads into my second point. When we project excessively, it leaves us empty, dead inside of ourselves. But also, it isn't just the bad parts of the self that are projected outward and *into* these creatures, *the robots themselves* are the projected bad parts of the self. That modern Prometheus, Frankenstein, provides a template for so many contemporary representations of robots: human endeavour, science and technology, from the best intentions, create nevertheless a monster, a creature that hubris leads us to believe that we can control. But the unnatural monster gains autonomy and cannot be submitted to our will. Our creation comes back to haunt us.

We see this story again and again in representations of robots. And like the monster in Mary Shelley's gothic horror, there is a warning here about reason. So many of our monsters since the nineteenth century – Frankenstein's creature, Mr. Hyde, Nazism, zombies and robots – are the terrible products of our own reason and our own science. H.A.L. 9000 [10], The Terminator, the Borg are ruthless in their efficiency; monsters made all the more destructive and potent by the fact that they are guided by a single principle – not an irrational violence, but a violence that is completely and utterly based in a calculated, indisputable *logic*, a fanatical dedication not to myth (as with the savage or the religious extremist) but to their technological, rational, scientific programming. Such monsters are the embodiment of our bad selves, the dreaded consequences of our reason, our science and our technology.

Does this mean we should fear robots now or in the future? No. Such cybernetic monsters are fictitious, meant to be object lessons, reminders that there are – or at least should be – limits on our mastery of the world through technology. However, I think that what makes these robots even more ultimately terrifying is the idea that they are the bad parts of ourselves that we know to fear – an unquestioning belief in science and an unbending devotion

to reason that depersonalises us, that makes *us* into the callous, in-humane monsters. If we project parts of ourselves we think bad, such as emotions, empathy or uncertainty – qualities that are integral to our humanity – in the quest to create ideal beings of reason, those empty, mechanical shells might well come back to destroy us.

To conclude, I want to introduce some preliminary remarks on the debate between the self-styled 'transhumanists' and those that they regard as their critics, whom they call 'bioconservatives'. I think this debate is instructive, and important, in the context of some of the issues I have raised here. The transhumanists – 'techno-enthusiast' thinkers such as Nick Bostrom, Aubrey de Grey, David Pearce and others – claim that humans and human nature are improvable 'through the use of applied science and other rational methods':

> Transhumanists imagine the possibilities in the near future of dramatically enhancing human mental and physical capacities, slowing and reversing the ageing process, and controlling our emotional and mental states. The imagined future is a new age in which people will be freed from mental disease and physical decrepitude, able to consciously choose their 'natures' and those of their children. [11]

Those, however, who oppose their aims, who are suspicious of the use of technology to modify humans and human nature, transhumanists label 'bioconservatives'. Some of these objections are based on religious grounds, while others object on the grounds of future inequality, or on Enlightenment humanist principles.

In the context of projection, we can see some basic differences between the two groups. Transhumanists, it seems, project good parts of the self into technology; in fact, some transhumanists hold out the possibility that one day perhaps the *entire* self – an entire consciousness – can be transferred, downloaded, into a machine, meaning that some *ideal self* will be projected *completely* into a technological container. The other group – who we will join the transhumanists for now in calling bioconservatives, though I don't think we can speak comfortably of them as a single group – see in technology a threat, the persecution of humanity's goodness. At some level, these thinkers seem to have in common a certain idealisation of nature, or of a human nature that they want preserved and which the transhumanists' future technology threatens. For the transhumanists, technology is idealised, an all-good (leading to a future all good-self) wherein technology successfully contains and thus preserves the best of the human race and acts as its salvation. It seems to me, however, that some of those qualities they deem 'bad' are some of those very qualities that we – right now – regard as essential to human nature: the

uncertainty and vulnerability that accompanies aging, reproduction, pain and death. I say 'right now', because I regard the notion of trans-historical 'human nature' to be itself a construct, another creation of ours that will inevitably change in the future, just as it has done in the past. It is a fantasy to regard any such conception as 'ideal' or 'inalienable', though how we idealise – or demonise – such conceptions says a great deal about the values that we wish to project.

Who is correct, the transhumanists or the bioconservatives? Neither, of course. For all projections are fantasies, based on part-objects, half-truths, wishful thinking and, at least on some level, paranoia – an irrational fear of one thing or another. It is only when we develop a greater *ambivalence* (by which I do not mean 'indifference' but an ability to balance bad and good in a sensible, balanced way) that we can engage with any object, including the robot, the idea of technology or our own technological prowess in a realistic, useful way. What we need to realise is that both groups' projections are based in fantasies, and it is those fantasies that must be explored in more depth. Projections are, in the beginning, at their heart and certainly at their most potent, ways in which we cope with anxiety, fantasies that we deploy to protect ourselves from badness. So the questions that need to be ask are, what fears lie behind the transhumanists' desires for the technologically-enhanced human? What anxieties lie behind the bioconservatives' resistance to this imagined future? Though these are questions for another study, it is only when we address these issues , I believe, that we will get to the core of this debate and understand what it is really about, the ground that each side is battling to defend, or the monsters that each is trying to keep at bay.

References

1. Dick, P. K.: Do Androids Dream of Electric Sheep? Kindle Edition. (1968)
2. Cameron, J.: The Terminator. Orion Pictures (1984)
3. Fontana, D. C., Roddenberry, G: Star Trek: The Next Generation (Encounter at Farpoint). Paramount Television (1987)
4. Park, N.: Wallace and Gromit: The Wrong Trousers. Aardman Animations (1993)
5. Klein, M.: Early Stages of the Oedipal Conflict. In: Love, Guilt and Reparation, and Other Works 1921-1945, pp. 186–198. Virago Press, London (1994)
6. Young, R.M.: Mental Space. Process Press, London (1994) `http://www.human-nature.com/rmyoung/papers/paper55.html`
7. Winnicott, D. W.: Through Paediatrics to Psychoanalysis: Collected Papers. Karnac Books, London (1958)

8. Bollas, C.: The Shadow of the Object: Psychoanalysis of the Unthought Known. Columbia University Press, New York (1987)
9. Kohon, G. (ed.): The Dead Mother: The Work of André Green. Routledge, New York (1999)
10. Kubrick, S.: 2001: A Space Odyssey. MGM (1968)
11. Hansell G.R., Grassie, W. (eds.): H+-: Transhumanism and its Critics. Philadelphia: Metanexus Institute. Kindle Edition (2011)

How We're Predicting AI – or Failing to

Stuart Armstrong[1] and Kaj Sotala[2]

[1] The Future of Humanity Institute,
Faculty of Philosophy, University of Oxford, Oxford, UK
stuart.armstrong@stx.oxon.org
[2] The Singularity Institute, Berkeley, CA, USA
kaj@singularity.org

Abstract. This paper will look at the various predictions that have been made about AI and propose decomposition schemas for analysing them. It will propose a variety of theoretical tools for analysing, judging and improving these predictions. Focusing specifically on timeline predictions (dates given by which we should expect the creation of AI), it will show that there are strong theoretical grounds to expect predictions to be quite poor in this area. Using a database of 95 AI timeline predictions, it will show that these expectations are born out in practice: expert predictions contradict each other considerably, and are indistinguishable from non-expert predictions and past failed predictions. Predictions that AI lie 15 to 25 years in the future are the most common, from experts and non-experts alike.

Keywords: AI, predictions, experts, bias

1 Introduction

Predictions about the future development of artificial intelligence are as confident as they are diverse. Starting with Turing's initial estimation of a 30% pass rate on Turing test by the year 2000 [1], computer scientists, philosophers and journalists have never been shy to offer their own definite prognostics, claiming AI to be impossible [2] or just around the corner [3] or anything in between.

What are we to make of these predictions? What are they for, and what can we gain from them? Are they to be treated as light entertainment, the equivalent of fact-free editorials about the moral decline of modern living? Or are there some useful truths to be extracted? Can we feel confident that certain categories of experts can be identified, and that their predictions stand out from the rest in terms of reliability? In this paper, we start off by proposing

classification schemes for AI predictions: what types of predictions are being made, and what kinds of arguments or models are being used to justify them.

Different models and predictions can result in very different performances, and it will be the ultimate aim of this project to classify and analyse their varying reliability. Armed with this scheme, we then analyse some of these approaches from the theoretical perspective, seeing whether there are good reasons to believe or disbelieve their results. The aim is not simply to critique individual methods or individuals, but to construct a toolbox of assessment tools that will both enable us to estimate the reliability of a prediction, and allow predictors to come up with better results themselves.

This paper, the first in the project, looks specifically at AI timeline predictions: those predictions that give a date by which we should expect to see an actual AI being developed (we use AI in the old fashioned sense of a machine capable of human-comparable cognitive performance; a less ambiguous modern term would be 'AGI', Artificial *General* Intelligence). With the aid of the biases literature, we demonstrate that there are strong reasons to expect that experts would *not* be showing particular skill in the field of AI timeline predictions. The task is simply not suited for good expert performance.

Those theoretical results are supplemented with the real meat of the paper: a database of 257 AI predictions, made in a period spanning from the 1950s to the present day. This database was assembled by researchers from the Singularity Institute (Jonathan Wang and Brian Potter) systematically searching though the literature, and is a treasure-trove of interesting results. A total of 95 of these can be considered AI timeline predictions. We assign to each of them a single 'median AI' date, which then allows us to demonstrate that AI expert predictions are greatly inconsistent with each other – and indistinguishable from non-expert performance, and past failed predictions.

With the data, we further test two folk theorems: firstly that predictors always predict the arrival of AI just before their own deaths, and secondly that AI is always 15 to 25 years into the future. We find evidence for the second thesis but not for the first. This enabled us to show that there seems to be no such thing as an "AI expert" for timeline predictions: no category of predictors stands out from the crowd.

2 Taxonomy of Predictions

2.1 Prediction Types

There will never be a bigger plane built.
Boeing engineer on the 247 (a twin engine plane that held ten people)

The standard image of a prediction is some fortune teller staring deeply into the mists of a crystal ball, and decreeing, with a hideous certainty, the course of the times to come. Or in a more modern version, a scientist predicting the outcome of an experiment or an economist pronouncing on next year's GDP figures. But these "at date X, Y will happen" are just one type of valid prediction. In general, a prediction is something that constrains our expectation of the future. Before hearing the prediction, we though the future would have certain properties; but after hearing and believing it, we now expect the future to be different from our initial thoughts.

Under this definition, conditional predictions – "if A, then B will happen" – are also perfectly valid. As are negative predictions: we might have believed initially that perpetual motion machines were possible, and imagined what they could be used for. But once we accept that one cannot violate conservation of energy, we have a different picture of the future: one without these wonderful machines and all their fabulous consequences.

For the present analysis, we will divide predictions about AI into four types:

1. Timelines and outcome predictions. These are the traditional types of predictions, telling us when we will achieve specific AI milestones. Examples: An AI will pass the Turing test by 2000 [1]; Within a decade, AIs will be replacing scientists and other thinking professions [4].
2. Scenarios. These are a type of conditional predictions, claiming that if the conditions of the scenario are met, then certain types of outcomes will follow. Example: If we build a human-level AI that is easy to copy and cheap to run, this will cause mass unemployment among ordinary humans [5].
3. Plans. These are a specific type of conditional prediction, claiming that if someone decides to implement a specific plan, then they will be successful in achieving a particular goal. Example: We can build an AI by scanning a human brain and simulating the scan on a computer [6].
4. Issues and metastatements. This category covers relevant problems with (some or all) approaches to AI (including sheer impossibility results), and metastatements about the whole field. Examples: an AI cannot be built

without a fundamental new understanding of epistemology [7]; Generic AIs will have certain (potentially dangerous) behaviours [8].

There will inevitably be some overlap between the categories, but this division is natural enough for our purposes. In this paper we will be looking at timeline predictions. Thanks to the efforts of Jonathan Wang and Brian Potter at the Singularity Institute, the authors were able to make use of extensive databases of this type of predictions, reaching back from the present day back to the 1950s. Other types of predictions will be analysed in subsequent papers.

2.2 Prediction Methods

Just as there are many types of predictions, there are many ways of arriving at them – consulting crystal balls, listening to the pronouncements of experts, constructing elaborate models. Our review of published predictions has shown that the prediction methods are far more varied than the types of conclusions arrived at. For the purposes of this analysis, we'll divide the prediction methods into the following loose scheme:

1. Causal models
2. Non-causal models
3. The outside view
4. Philosophical arguments
5. Expert authority
6. Non-expert authority

Causal model are the staple of physics: given certain facts about the situation under consideration (momentum, energy, charge, etc...) a conclusion is reached about what the ultimate state will be. If the facts were different, the end situation would be different.

But causal models are often a luxury outside of the hard sciences, whenever we lack precise understanding of the underlying causes. Some success can be achieved with non-causal models: without understanding what influences what, one can extrapolate trends into the future. Moore's law is a highly successful non-causal model [9].

The outside view is a method of predicting that works by gathering together specific examples and claiming that they all follow the same underlying trend. For instance, one could notice the plethora of Moore's laws across the spectrum of computing (in numbers of transistors, size of hard drives, network capacity, pixels per dollar...), note that AI is in the same category, and hence argue that AI development must follow a similarly exponential curve [10].

Philosophical arguments are common in the field of AI; some are simple impossibility statements: AI is decreed to be impossible for more or less plausible reasons. But the more thoughtful philosophical arguments point out problems that need to be resolved to achieve AI, highlight interesting approaches to doing so, and point potential issues if this were to be achieved.

Many predictions rely strongly on the status of the predictor: their innate expertise giving them potential insights that cannot be fully captured in their arguments, so we have to trust their judgment. But there are problems in relying on expert opinion, as we shall see.

Finally, some predictions rely on the judgment or opinion of non-experts. Journalists and authors are examples of this, but often actual experts will make claims outside their domain of expertise. CEO's, historians, physicists and mathematicians will generally be no more accurate than anyone else when talking about AI, no matter how stellar they are in their own field [11].

Predictions can use a mixture of these approaches, and often do. For instance, Ray Kurzweil's 'Law of Time and Chaos' uses the outside view to group together evolutionary development, technological development, and computing into the same category, and constructs a causal model predicting time to the 'Singularity' [10]. Moore's law (non-causal model) is a key input to this Law, and Ray Kurzweil's expertise is the main evidence for the Law's accuracy.

This is the schema we will be using in this paper, and in the prediction databases we have assembled. But the purpose of any such schema is to bring clarity to the analysis, not to force every prediction into a particular box. We hope that the methods and approaches used in this paper will be of general use to everyone wishing to analyse the reliability and usefulness of predictions, in AI and beyond. Hence this schema can be freely adapted or discarded if a particular prediction does not seem to fit it, or if an alternative schema seems to be more useful for the analysis of the question under consideration.

3 A Toolbox of Assessment Methods

The purpose of this paper is not only to assess the accuracy and reliability of some of the AI predictions that have already been made. The purpose is to start building a 'toolbox' of assessment methods that can be used more generally, applying them to current and future predictions.

3.1 Extracting Verifiable Predictions

The focus of this paper is squarely on the behaviour of AI. This is not a philosophical point; we are not making the logical positivist argument that only empirically verifiable predictions have meaning [12]. But it must be noted that many of the vital questions about AI – can it built, when, will it be dangerous, will it replace humans, and so on – all touch upon behaviour. This narrow focus has the added advantage that empirically verifiable predictions are (in theory) susceptible to falsification, which means ultimately agreement between people of opposite opinions. Predictions like these have a very different dynamic to those that cannot be shown to be wrong, even in principle.

To that end, we will seek to reduce the prediction to an empirically verifiable format. For some predictions, this is automatic: they are already in the correct format. When Kurzweil wrote "One of my key (and consistent) predictions is that a computer will pass the Turing test by 2029," then there is no need to change anything. Conversely, some philosophical arguments concerning AI, such as some of the variants of the Chinese Room argument [13], are argued to contain no verifiable predictions at all: an AI that demonstrated perfect human behaviour would not affect the validity of the argument.

And in between there are those predictions that are partially verifiable. Then the verifiable piece must be clearly extracted and articulated. Sometimes it is ambiguity that must be overcome: when an author predicts an AI "Omega point" in 2040 [14], it is necessary to read the paper with care to figure out what counts as an Omega point and (even more importantly) what doesn't.

Even purely philosophical predictions can have (or can be interpreted to have) verifiable predictions. One of the most famous papers on the existence of conscious states is Thomas Nagel's "What is it like to be a bat." [15]. In this paper, Nagel argues that bats must have mental states, but that we humans can never understand what it is like to have these mental states. This feels purely philosophical, but does lead to empirical predictions: that if the bat's intelligence were increased and we could develop a common language, then at some point in the conversation with it, our understanding would reach an impasse. We would try to describe what our internal mental states felt like, but would always fail to communicate the essence of our experience to the other species.

Many other philosophical papers can likewise be read as having empirical predictions; as making certain states of the world more likely or less – even if they seem to be devoid of this. The Chinese Room argument, for instance, argues that formal algorithms will lack the consciousness that humans possess [13]. This may seem to be an entirely self-contained argument – but consider

that a lot of human behaviour revolves around consciousness, be it discussing it, commenting on it, defining it or intuitively noticing it in others. Hence if we believed the Chinese Room argument, and were confronted with two AI projects, one based on advanced algorithms and one based on modified human brains, we would be likely to believe that the second project is more likely to result in an intelligence that *seemed* conscious than the first. This is simply because we wouldn't believe that the first AI could ever be conscious, and that it is easier to seem conscious when one actually is. And that gives an empirical prediction.

Note that the authors of the predictions may disagree with our 'extracted' conclusions. This is not necessarily a game breaker. For instance, even if there is no formal link between the Chinese Room model and the prediction above, it's still the case that the intuitive reasons for believing the model are also good reasons for believing the prediction. Our aim should always be to try and create useful verifiable predictions in any way we can. In this way, we can make use of much more of the AI literature. For instance, Lucas argues that AI is impossible because it could not recognise the truth of its own Gödel sentence[3][16]. This is a very strong conclusion, and we have to accept a lot of Lucas's judgments before we agree with it. Replacing the conclusion with the weaker (and verifiable) "self reference will be an issue with advanced AI, and will have to be dealt with somehow by the programmers" gives us a useful prediction which is more likely to be true.

Care must be taken when applying this method: the point is to extract a useful verifiable prediction, not to weaken or strengthen a reviled or favoured argument. The very first stratagems in Shopenhauer's "The Art of Always being Right" [17] are to extend and over-generalise the consequences of your opponent's argument; conversely, one should reduce and narrow down one's own arguments. There is no lack of rhetorical tricks to uphold one's own position, but if one is truly after the truth, one must simply attempt to find the most reasonable empirical version of the argument; the truth-testing will come later.

This method often increases uncertainty, in that it often narrows the consequences of the prediction, and allows more possible futures to exist, consistently with that prediction. For instance, Bruce Edmonds [18], building on the "No Free Lunch" results [19], demonstrates that there is no such thing as

[3] A Gödel sentence is a sentence G that can be built in any formal system containing arithmetic. G is implicitly self-referential, as it is equivalent with "there cannot exist a proof of G". By construction, there cannot be a consistent proof of G from within the system.

a universal intelligence: no intelligence that performs better than average in every circumstance. Initially this seems to rule out AI entirely; but when one analyses what this means empirically, one realises there is far less to it. It does not forbid an algorithm from performing better than any human being in any situation any human being would ever encounter, for instance. So our initial intuition, which was to rule out all futures with AIs in them, is now replaced by the realisation that we have barely put any constraints on the future at all.

3.2 Clarifying and Revealing Assumptions

The previous section was concerned with the predictions' conclusions. Here we will instead be looking at its assumptions, and the logical structure of the argument or model behind it. The objective is to make the prediction as rigorous as possible

Philosophers love doing this: taking apart argument, adding caveats and straightening out the hand-wavy logical leaps. In a certain sense, it can be argued that analytic philosophy is entirely about making arguments rigorous. One of the oldest methods in philosophy – the dialectic [20] – also plays this role, with concepts getting clarified during the conversation between philosophers and various Athenians. Though this is perhaps philosophy's greatest contribution to knowledge, it is not exclusively the hunting ground of philosophers. All rational fields of endeavour do – and should! – benefit from this kind of analysis.

Of critical importance is revealing hidden assumptions that went into the predictions. These hidden assumptions – sometimes called Enthymematic gaps in the literature [21] – are very important because they clarify where the true disagreements lie, and where we need to focus our investigation in order to find out the truth of prediction. Too often, competing experts will make broad-based arguments that fly past each other. This makes choosing the right argument a matter of taste, prior opinions and our admiration of the experts involved. But if the argument can be correctly deconstructed, then the source of the disagreement can be isolated, and the issue can be decided on much narrower grounds – and its much clearer whether the various experts have relevant expertise or not (see Section 3.4). The hidden assumptions are often implicit, so it is perfectly permissible to construct assumptions that the predictors were not consciously aware of using.

For example, let's look again at the Gödel arguments mentioned in the Section 3.1. The argument shows that formal systems of a certain complexity must be either incomplete (unable to see that their Gödel sentence is true) or

inconsistent (proving false statements). This is contrasted with humans, who – allegedly – use meta-reasoning to know that their own Gödel statements are true. It should first be noted here that no one has written down an actual "human Gödel statement," so we cannot be sure humans would actually figure out that it is true[4]. Also, humans are both inconsistent and able to deal with inconsistencies without a complete collapse of logic. In this, they tend to differ from AI systems, though some logic systems such as relevance logic do mimic the same behaviour [22]. In contrast, both humans and AIs are not logically omniscient – they are not capable of proving everything provable within their logic system (the fact that there are an infinite number of things to prove being the problem here). So this analysis demonstrates the hidden assumption in Lucas's argument: that the behaviour of an actual computer program running on a real machine is more akin to that of a logically omniscient formal agent, than it would be to a real human being. That assumption may be flawed or correct, but is one of the real sources of disagreement over whether Gödelian arguments rule out artificial intelligence.

Again, it needs to be emphasised that the purpose is to clarify and analyse arguments, not to score points for one side or the other. It is easy to phrase assumptions in ways that sound good or bad for either "side". It is also easy to take the exercise too far: finding more and more minor clarifications or specific hidden assumptions until the whole prediction becomes a hundred page mess of over-detailed special cases. The purpose is to clarify the argument until it reaches the point where all (or most) parties could agree that these assumptions are the real sources of disagreement. And then we can consider what empirical evidence, if available, or expert opinion has to say about these disagreements.

There is surprisingly little published on the proper way of clarifying assumptions, making this approach more an art than a science. If the prediction comes from a model, we have some standard tools available for clarifying, though [23]. Most of these methods work by varying parameters in the model and checking that this doesn't cause a breakdown in the prediction.

Model Testing and Counterfactual Resiliency Though the above works from inside the model, there are very few methods that can test the strength of a model from the outside. This is especially the case for non-causal models: what are the assumptions behind Moore's famous law [9], or Robin Hanson's model that we are due for another technological revolution, based on the

[4] One could argue that, by definition, a human Gödel statement must be one that humans cannot recognise as being a human Gödel statement!

timeline of previous revolutions [24]? If we can't extract assumptions, we're reduced to saying "that feel right/wrong to me", and therefore we're getting nowhere.

The authors have come up with a putative way of testing the assumptions of such models (in the case of Moore's law, the empirical evidence in favour is strong, but there is still the question of what is powering the law and whether it will cross over to new chip technologies again and again). It involves giving the model a counterfactual resiliency check: imagining that world history had happened slightly differently, and checking whether the model would have stood up in those circumstances. Counterfactual changes are permitted to anything that the model ignores.

The purpose of this exercise is not to rule out certain models depending on one's own preferred understanding of history (e.g. "Protestantism was essential to the industrial revolution, and was a fluke due to Martin Luther; so it's very likely that the industrial revolution would not have happened in the way or timeframe that it did, hence Hanson's model – which posits the industrial revolutions's dates as inevitable – is wrong"). Instead it is to illustrate the tension between the given model and other models of history (e.g. "The assumptions that Protestantism was both a fluke and essential to the industrial revolution are in contradiction with Hanson's model. Hence Hanson's model implies that either Protestantism was inevitable or that it was non-essential to the industrial revolution, a extra hidden assumption"). The counterfactual resiliency exercise has been carried out at length in an online post[5]. The general verdict seemed to be that Hanson's model contradicted a lot of seemingly plausible assumptions about technological and social development. Moore's law, on the other hand, seemed mainly dependent on the continuing existence of a market economy and the absence of major catastrophes.

This method is new, and will certainly be refined in future. Again, the purpose of the method is not to rule out certain models, but to find the nodes of disagreement.

More Uncertainty Clarifying assumptions often ends up increasing uncertainty, as does revealing hidden assumptions. The previous section focused on extracting verifiable predictions, which often increases the range of possible worlds compatible with a prediction. Here, by clarifying and caveatting assumptions, and revealing hidden assumption, we reduce the number of worlds in which the prediction is valid. This means that the prediction puts fewer

[5] See `http://lesswrong.com/lw/ea8/counterfactual_resiliency_test_for_noncausal`

constraints on our expectations. In counterpart, of course, the caveatted prediction is more likely to be true.

3.3 Empirical Evidence

The gold standard in separating true predictions from false ones must always be empirical evidence. The scientific method has proved to be the best way of disproving false hypotheses, and should be used whenever possible. Other methods, such as expert opinion or unjustified models, come nowhere close.

The problem with empirical evidence is that... it is generally non-existent in the AI prediction field. Since AI predictions are all about the existence and properties of a machine that hasn't yet been built, that no-one knows how to build or whether it actually can be built, there is little opportunity for the whole hypothesis-prediction-testing cycle. This should indicate the great difficulties in the field. Social sciences, for instance, are often seen as the weaker cousins of the hard sciences, with predictions much more contentious and less reliable. And yet the social sciences make use of the scientific method, and have access to some types of repeatable experiments. Thus any prediction in the field of AI should be treated as less likely than any social science prediction.

That generalisation is somewhat over-harsh. Some AI prediction methods hew closer to the scientific method, such as the whole brain emulations model [6] – it makes testable predictions along the way. Moore's law is a wildly successful prediction, and connected to some extent with AI. Many predictors (e.g. Kurzweil) make partial predictions on the road towards AI; these can and should be assessed – track records allow us to give some evidence to the proposition "this expert knows what they're talking about." And some models also allow for a degree of testing. So the field is not void of empirical evidence; it's just that there is so little of it, and to a large extent we must put our trust in expert opinion.

3.4 Expert Opinion

Reliance on experts is nearly unavoidable in AI prediction. Timeline predictions are often explicitly based on experts' feelings; even those that consider factors about the world (such as computer speed) need an expert judgment about why that factor is considered and not others. Plans need experts to come up with them and judge their credibility. And unless every philosopher agrees on the correctness of a particular philosophical argument, we are dependent to some degree on the philosophical judgment of the author. It is the

purpose of all the methods described above that we can refine and caveat a prediction, back it up with empirical evidence whenever possible, and thus clearly highlight the points where we need to rely on expert opinion. And so can focus on the last remaining points of disagreement: the premises themselves (that is of course the ideal situation: some predictions are given directly with no other basis but expert authority, meaning there is nothing to refine).

Should we expect experts to be good at this task? There have been several projects over the last few decades to establish the domains and tasks where we would expect experts to have good performance [25, 26]. Table 1 summarises the results:

Table 1. Table of task properties conducive to good and poor expert performance.

Good performance:	Poor performance:
Static stimuli	Dynamic (changeable) stimuli
Decisions about things	Decisions about behaviour
Experts agree on stimuli	Experts disagree on stimuli
More predictable problems	Less predictable problems
Some errors expected	Few errors expected
Repetitive tasks	Unique tasks
Feedback available	Feedback unavailable
Objective analysis available	Subjective analysis only
Problem decomposable	Problem not decomposable
Decision aids common	Decision aids rare

Not all of these are directly applicable to the current paper (are predictions about human level AIs predictions about things, or about behaviour?). One of the most important factors is whether experts get feedback, preferably immediate feedback. We should expect the best expert performance when their guesses are immediately confirmed or disconfirmed. When feedback is unavailable or delayed, or the environment isn't one that give good feedback, then expert performance drops precipitously [26, 11].

Table 1 applies to both domain and task. Any domain of expertise strongly in the right column will be one where we expect poor expert performance. But if the individual expert tries to move their own predictions into the left column (maybe by decomposing the problem as far as it will go, training themselves on related tasks where feedback is available. . .) they will be expected to perform better. In general, we should encourage this type of approach.

When experts fail, there are often simple algorithmic models that demonstrate better performance [27]. In these cases, the experts often just spell out their criteria, design the model in consequence, and let the model give its predictions: this results in better predictions than simply asking the expert in the first place. Hence we should also be on the lookout for experts who present their findings in the form of a model.

As everyone knows, experts sometimes disagree. This fact strikes at the very heart of their supposed expertise. We listen to them because they have the skills and experience to develop correct insights. If other experts have gone through the same process and come to an opposite conclusion, then we have to conclude that their insights do not derive from their skills and experience, and hence should be discounted. Now if one expert opinion is a fringe position held by only a few experts, we may be justified in dismissing it simply as an error. But if there are different positions held by large numbers of disagreeing experts, how are we to decide between them? We need some sort of objective criteria: we are not experts in choosing between experts, so we have no special skills in deciding the truths on these sorts of controversial positions.

What kind of objective criteria could there be? A good track record can be an indicator, as is a willingness to make verifiable, non-ambiguous predictions. A better connection with empirical knowledge and less theoretical rigidity are also positive indications [28], and any expert that approached their task with methods that were more on the left of the table than on the right should be expected to be more correct. But these are second order phenomena – we're looking at our subjective interpretation of expert's subjective opinion – so in most cases, when there are strong disagreement between experts, we simply can't tell which position is true.

Grind versus Insight Some AI prediction claim that AI will result from grind: i.e. lots of hard work and money. Other claim that AI will need special insights: new unexpected ideas that will blow the field wide open [7].

In general, we are quite good at predicting grind. Project managers and various leaders are often quite good at estimating the length of projects (as long as they're not directly involved in the project [29]). Even for relatively creative work, people have sufficient feedback to hazard reasonable guesses. Publication dates for video games, for instance, though often over-optimistic, are generally not ridiculously erroneous – even though video games involve a lot of creative design, play-testing, art, programing the game "AI", etc... Moore's law could be taken as an ultimate example of grid: we expect the global efforts

of many engineers across many fields to average out to a rather predictable exponential growth.

Predicting insight, on the other hand, seems a much more daunting task. Take the Riemann hypothesis, a well-established mathematical hypothesis from 1885, [30]. How would one go about estimating how long it would take to solve? How about the $P = NP$ hypothesis in computing? Mathematicians seldom try and predict when major problems will be solved, because they recognise that insight is very hard to predict. And even if predictions could be attempted (the age of the Riemann's hypothesis hints that it probably isn't right on the cusp of being solved), they would need much larger error bars than grind predictions. If AI requires insights, we are also handicapped by the fact of not knowing what these insights are (unlike the Riemann hypothesis, where the hypothesis is clearly stated, and only the proof is missing). This could be mitigated somewhat if we assumed there were several different insights, each of which could separately lead to AI. But we would need good grounds to assume that.

Does this mean that in general predictions that are modeling grind should be accepted more than predictions that are modeling insight? Not at all. Predictions that are modeling grind should only be accepted if they can make a good case that producing an AI is a matter grind only. The predictions around whole brain emulations [6], are one of the few that make this case convincingly; this will be analysed in a subsequent paper.

Non-Experts Opinion It should be born in mind that all the caveats and problems with non-expert opinion apply just as well to non-experts. With one crucial difference: we have no reason to trust the non-expert's opinion in the first place. That is not to say that non-experts cannot come up with good models, convincing timelines, or interesting plans and scenarios. It just means that our assessment of the quality of the prediction depends only on what we are given; we cannot extend a non-expert any leeway to cover up a weak premise or a faulty logical step. To ensure this, we should try and assess non-expert predictions blind, without knowing who the author is. If we can't blind them, we can try and get a similar effect by asking ourselves hypothetical questions such as: "Would I find this prediction more or less convincing if the author was the Archbishop of Canterbury? What if it was Warren Buffet? Or the Unabomber?" We should aim to reach the point where hypothetical changes in authorship do not affect our estimation of the prediction.

4 Timeline Predictions

The practical focus of this paper is on AI timeline predictions: predictions giving dates for AIs with human-comparable cognitive abilities. Researchers from the Singularity Institute have assembled a database of 257 AI predictions since 1950, of which 95 include AI timelines.

4.1 Subjective Assessment

A brief glance at Table 1 allows us to expect that AI timeline predictions will generally be of very poor quality. The only factor that is unambiguously positive for AI predictions is that prediction errors are expected and allowed: apart from that, the task seems singularly difficult, especially on the key issue of feedback. An artificial intelligence is a hypothetical machine, which has never existed on this planet before and about whose properties we have but the haziest impression. Most AI experts will receive no feedback whatsoever about their predictions, meaning they have to construct them entirely based on their untested impressions.

There is nothing stopping experts from decomposing the problem, or constructing models which they then calibrate with available data, or putting up interim predictions to test their assessment. And some do use these better approaches (see for instance [10, 5, 31]). But a surprisingly large number don't! Some predictions are unabashedly based simply on the feelings of the predictor [32, 33].

Yet another category are of the "Moore's law hence AI" type. They postulate that AI will happen when computers reach some key level, often comparing with some key property of the brain (number of operations per second [34], or neurones/synapses[6]). In the division established in section 3.4, this is pure 'grind' argument: AI will happen after a certain amount of work is performed. But, as we saw, these kinds of arguments are only valid if the predictor has shown that reaching AI does not require new insights! And that step is often absent from the argument.

4.2 Timeline Prediction Data

The above were subjective impressions, formed while looking over the whole database. To enable more rigorous analysis, the various timeline predictions

[6] See for instance Dani Eder's 1994 Newgroup posting `http://www.aleph.se/Trans/Global/Singularity/singul.txt`

were reduced to a single number for purposes of comparison: this would be the date upon which the predictor expected 'human level AI' to be developed.

Unfortunately not all the predictions were in the same format. Some gave ranges, some gave median estimates, some talked about superintelligent AI, others about slightly below-human AI. In order to make the numbers comparable, one of the authors (Stuart Armstrong) went through the list and reduced the various estimates to a single number. He followed the following procedure to extract a "Median human-level AI estimate":

When a range was given, he took the mid-point of that range (rounded down). If a year was given with a 50% likelihood estimate, he took that year. If it was the collection of a variety of expert opinions, he took the prediction of the median expert. If the predictor forsaw some sort of AI by a given date (partial AI or superintelligent AI), and gave no other estimate, he took that date as their estimate rather than trying to correct it in one direction or the other (there were roughly the same number of subhuman AIs as suphuman AIs in the list, and not that many of either). He read extracts of the papers to make judgement calls when interpreting problematic statements like "within thirty years" or "during this century" (is that a range or an end-date?). Every date selected was either an actual date given by the predictor, or the midpoint of a range.[7]

It was also useful to distinguish between popular estimates, performed by journalists, writers or amateurs, from those predictions done by those with expertise in relevant fields (AI research, computer software development, etc. . .) Thus each prediction was noted as 'expert' or 'non-expert'; the expectation being that experts would demonstrate improved performance over non-experts.

Figure 1 graphs the results of this exercise (the range has been reduced; there were seven predictions setting dates beyond the year 2100, three of them expert.)

As can be seen, expert predictions span the whole range of possibilities and seem to have little correlation with each other. The range is so wide – fifty year gaps between predictions are common – that it provides strong evidence that experts are not providing good predictions. There does not seem to be any visible difference between expert and non-expert performance either, suggesting that the same types of reasoning may be used in both situations, thus negating the point of expertise.

[7] The data can be found at `http://www.neweuropeancentury.org/SIAI-FHI_AI_predictions.xls`; readers are encouraged to come up with their own median estimates.

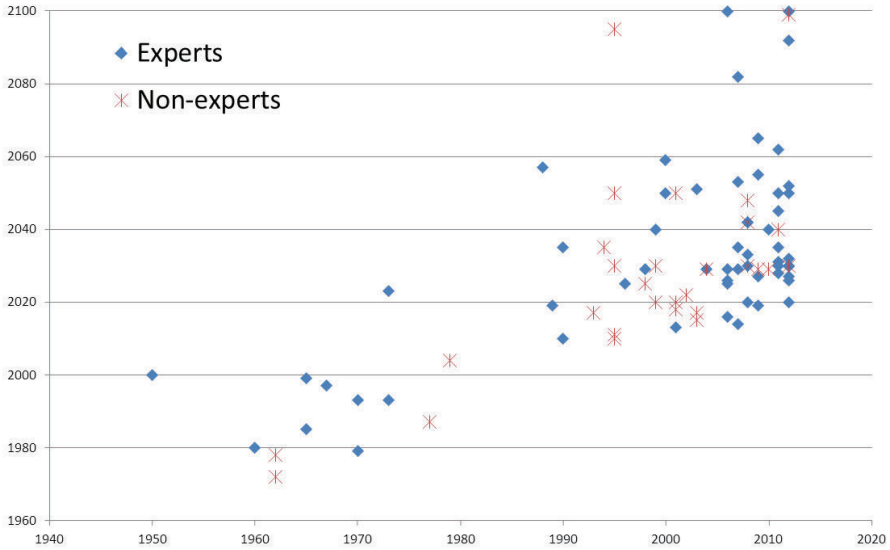

Fig. 1. Median estimate for human-level AI, graphed against date of prediction.

Two explanations have been generally advanced to explain poor expert performance in these matters. The first, the so-called Maes-Garreau law[8] posits that AI experts predict AI happening towards the end of their own lifetime. This would make AI into a technology that would save them from their own deaths, akin to a 'Rapture of the Nerds'.

The second explanation is that AI is perpetually fifteen to twenty-five years into the future. In this way (so the explanation goes), the predictor can gain credit for working on something that will be of relevance, but without any possibility that their prediction could be shown to be false within their current career. We'll now look at the evidence for these two explanations.

Nerds Don't Get Raptured Fifty-five predictions were retained, in which it was possible to estimate the predictor's expected lifespan. Then the difference between their median prediction and this lifespan was computed (a positive difference meaning they would expect to die before AI, a negative

[8] Kevin Kelly, editor of Wired magazine, created the law in 2007 after being influenced by Pattie Maes at MIT and Joel Garreau (author of Radical Evolution).

difference meaning they didn't). A zero difference would be a perfect example of the Maes-Garreau law: the predictor expects AI to be developed at the exact end of their life. This number was then plotted again the predictor's age in Figure 2 (the plot was restricted to those predictions within thirty years of the predictor's expected lifetime).

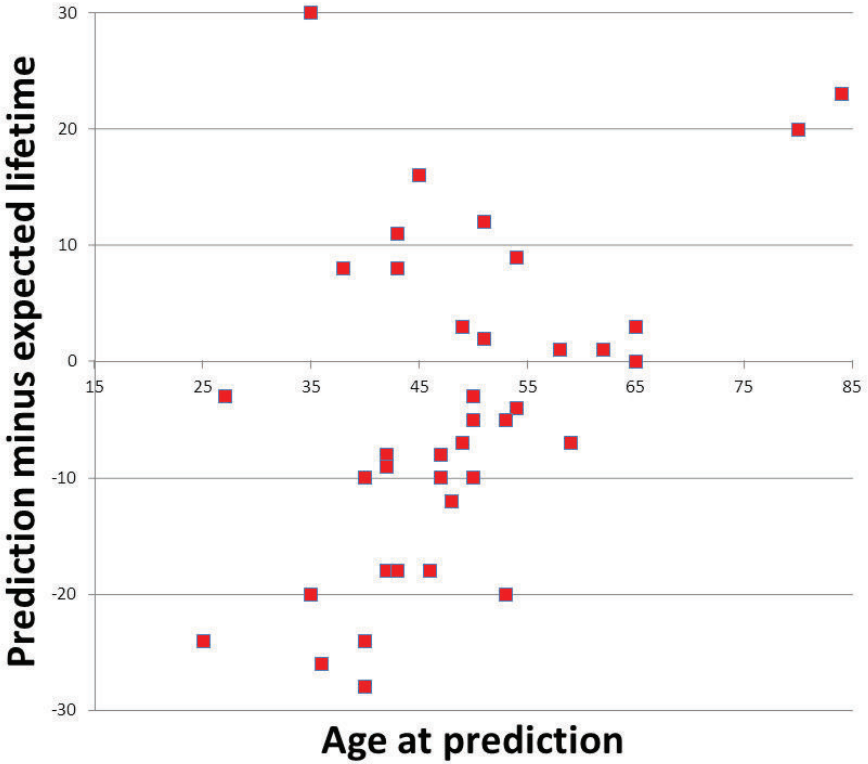

Fig. 2. Difference between the predicted time to AI and the predictor's life expectancy, graphed against the predictor's age.

From this, it can be seen that the Maes-Garreau law is not born out by the evidence: only twelve predictions (22% of the total) were within five years in either direction of the zero point.

Twenty Years to AI The 'time to AI' was computed for each expert prediction. This was graphed in Figure 3. This demonstrates a definite increase in the $16-25$ year predictions: 21 of the 62 expert predictions were in that range (34%). This can be considered weak evidence that experts do indeed prefer to predict AI happening in that range from their own time.

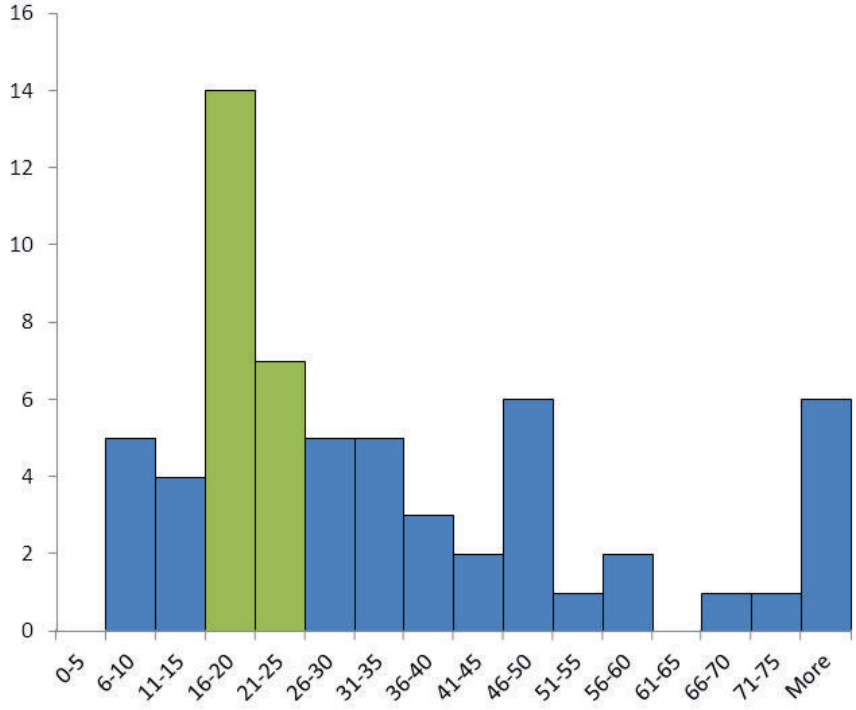

Fig. 3. Time between the arrival of AI and the date the prediction was made, for expert predictors.

But the picture gets more damning when we do the same plot for the non-experts, as in Figure 4. Here, 13 of the 33 predictions are in the $16-25$ year range. But more disturbingly, the time to AI graph is almost identical for experts and non-experts! Though this does not preclude the possibility of experts being more accurate, it does hint strongly that experts and non-

experts may be using similar psychological procedures when creating their estimates.

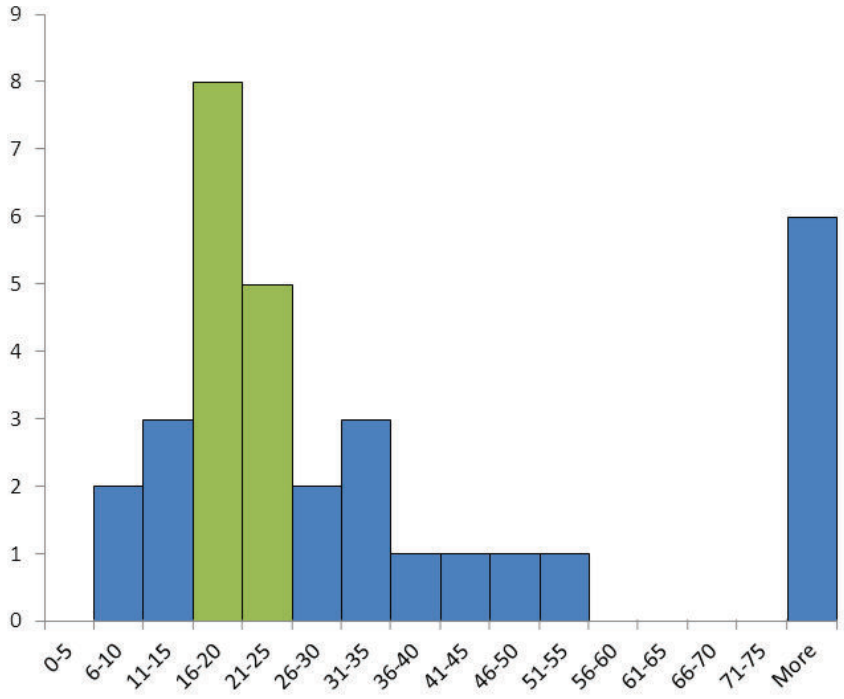

Fig. 4. Time between the arrival of AI and the date the prediction was made, for non-expert predictors.

The next step is to look at failed predictions. There are 15 of those, most dating to before the 'AI winter' in the eighties and nineties. These have been graphed in Figure 5 – and there is an uncanny similarity with the other two graphs! So expert predictions are not only indistinguishable from non-expert predictions, they are also indistinguishable from past failed predictions. Hence it is not unlikely that recent predictions are suffering from the same biases and errors as their predecessors

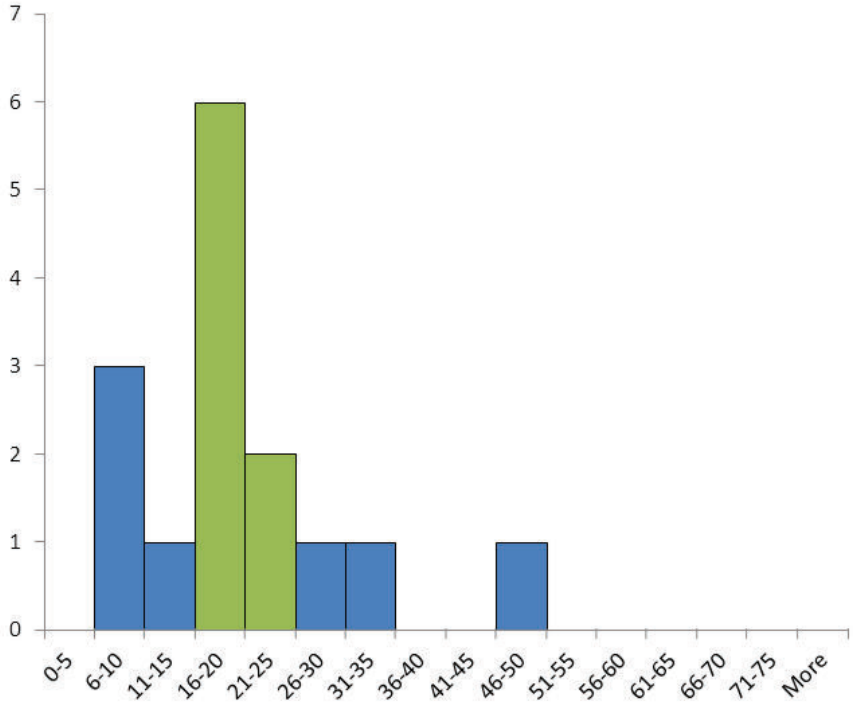

Fig. 5. Time between the arrival of AI and the date the prediction was made, for failed predictions.

5 Conclusion

This paper, the first in a series analysing AI predictions, focused on the reliability of AI timeline predictions (predicting the dates upon which 'human-level' AI would be developed). These predictions are almost wholly grounded on expert judgment. The biases literature classified the types of tasks on which experts would have good performance, and AI timeline predictions have all the hallmarks of tasks on which they would perform badly.

This was born out by the analysis of 95 timeline predictions in the database assembled by the Singularity Institute. There were strong indications therein that experts performed badly. Not only were expert predictions spread across a wide range and in strong disagreement with each other, but there was evidence that experts were systematically preferring a '15 to 25 years into the future' prediction. In this, they were indistinguishable from non-experts, and from past predictions that are known to have failed. There is thus no indication that experts brought any added value when it comes to estimating AI timelines. On the other hand, another theory – that experts were systematically predicting AI arrival just before the end of their own lifetime – was seen to be false in the data we have.

There is thus strong grounds for dramatically increasing the uncertainty in any AI timeline prediction.

Acknowledgments. The authors wish to acknowledge the help and support of the Singularity Institute, the Future of Humanity Institute and the James Martin School, as well as the individual advice of Nick Bostrom, Luke Muelhauser, Vincent Mueller, Anders Sandberg, Lisa Makros, Sean O'Heigeartaigh, Daniel Dewey, Eric Drexler and the online community of Less Wrong.

References

1. Turing, A.: Computing machinery and intelligence. Mind 59, 433–460(1950)
2. Jacquette, D.: Metamathematical criteria for minds and machines. Erkenntnis 27(1) (1987)
3. Darrach, B.: Meet shakey, the first electronic person. Reflections of the Future (1970)
4. Hall, J.: Further reflections on the timescale of AI. In: Solomonoff 85th Memorial Conference. (2011)
5. Hanson, R.: What if uploads come first: The crack of a future dawn. Extropy 6(2) (1994)

6. Sandberg, A.: Whole brain emulations: a roadmap. Future of Humanity Institute Technical Report 2008-3 (2008)
7. Deutsch, D.: The very laws of physics imply that artificial intelligence must be possible. What's holding us up? Aeon (2012)
8. Omohundro, S.: Basic AI drives. In: Proceedings of the First AGI Conference. Volume 171. (2008)
9. Moore, G.: Cramming more components onto integrated circuits. Electronics 38(8) (1965)
10. Kurzweil, R.: The Age of Spiritual Machines: When Computers Exceed Human Intelligence. Viking Adult (1999)
11. Kahneman, D.: Thinking, Fast and Slow. Farra, Straus and Giroux (2011)
12. Carnap, R.: The Logical Structure of the World. (1928)
13. Searle, J.: Minds, brains and programs. Behavioral and Brain Sciences 3(3), 417–457 (1980)
14. Schmidhuber, J. The New AI: General & Sound & Relevant for Physics In: Goertzel, B., Pennachin, C. (eds.) Artificial General Intelligence, pp. 177–200 (2006)
15. Nagel, T.: What is it like to be a bat? The Philosophical Review 83(4), 435–450 (1974)
16. Lucas, J.: Minds, machines and Gödel. Philosophy XXXVI, 112–127 (1961)
17. Schopenhauer, A.: The Art of Being Right: 38 Ways to Win an Argument. (1831)
18. Edmonds, B.: The social embedding of intelligence. In: Parsing the Turing Test, pp. 211–235. Springer Netherlands (2009)
19. Wolpert, D., Macready, W.: No free lunch theorems for search. (1995)
20. Plato: The Republic. (380 BC)
21. Fallis, D.: Intentional gaps in mathematical proofs. Synthese 134(1-2) (2003)
22. Routley, R., Meyer, R.: Dialectical logic, classical logic, and the consistency of the world. Studies in East European Thought 16(1-2) (1976)
23. Morgan, M., Henrion, M.: Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis. Cambridge University Press (1990)
24. Hanson, R.: Economics of brain emulations. In: Unnatrual Selection - The Challenges of Engineering Tomorrow's People, pp. 150–158 (2008)
25. Shanteau, J.: Competence in experts: The role of task characteristics. Organizational Behavior and Human Decision Processes 53, 252–266 (1992)
26. Kahneman, D., Klein, G.: Conditions for intuitive expertise: A failure to disagree. American Psychologist 64(6), 515–526 (2009)
27. Grove, W., Zald, D., Lebow, B., Snitz, B., Nelson, C.: Clinical versus mechanical prediction: A meta-analysis. Psychological Assessment 12, 19–30 (2000)
28. Tetlock, P.: Expert political judgement: How good is it? How can we know? (2005)
29. Buehler, R., Griffin, D., Ross, M.: Exploring the planning fallacy: Why people underestimate their task completion times. Journal of Personality and Social Psychology 67, 366–381 (1994)

30. Riemann, B.: Ueber die anzahl der primzahlen unter einer gegebenen grösse. Monatsberichte der Berliner Akademie (1859)
31. Waltz, D.: The prospects for building truly intelligent machines. Daedalus 117(1) (1988)
32. Good, J.: The scientist speculates: an anthology of partly-baked ideas. Heinemann (1962)
33. Armstrong, S.: Chaining god: A qualitative approach to AI, trust and moral systems. (2007) Online article.
34. Bostrom, N.: How long before superintelligence? International Journal of Futures Studies 2 (1998)

Is There Something Beyond AI? Frequently Emerging, but Seldom Answered Questions about Artificial Super-Intelligence

Jiří Wiedermann

Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, Czech Republic
jiri.wiedermann@cs.cas.cz

Abstract. Using the contemporary theories and views of computing and of cognitive systems we indicate possible answers to the following frequently asked questions about artificial intelligence: (i) what is the "computational power" of artificial cognitive systems?; (ii) are there "levels" of intelligence?; (iii) what is the position of human intelligence w.r.t. the "levels" of intelligence?; (iv) is there a general mechanism of intelligence?; (v) can "fully-fledged" body-less intelligence exist?; (vi) can there exist a sentient cloud? We give qualified arguments suggesting that within the large class of computational models of cognitive systems the answers to the previous question are positive. These arguments are mostly based on the author's recent works related to this problematics.

Keywords: cognitive systems, computional models, non-uniform evolving automaton

1 Introduction

Let us consider the question, whether there is something beyond artificial intelligence. A possible reading of this question is, whether there is a kind of intelligence that is, in some sense, more powerful that any kind of artificial intelligence.

In order to answer such a general question we break it down into a number of related, more concrete sub-questions: (i) what is the "computational power" of artificial cognitive systems?; (ii) are there "levels" of intelligence?; (iii) what is the position of human intelligence w.r.t. the "levels" of intelligence?; (iv) is there a general mechanism of intelligence?; (v) can "fully-fledged" body-less intelligence exist? and, last but not least, (vi) can there exist a sentient cloud?

Looking for the respective answers, our quest will be based on computational models of cognitive systems. This is because computational models present a truly large class of systems and until now no cognitive mechanisms in natural cognitive systems (living organisms) have been identified that could not be modelled computationally.

Our arguments will be based on four computational models co-authored by the present author in recent times each of which captures a different aspect of computational cognitive systems. In the next section we will consider each of the previous questions separately, giving the respective answers using appropriate computational model capturing the essence of the questions.

2 Answering the Questions

In order to give qualified answers to our questions we will refer to various non-standard computational (or algorithmic) models of general computational or specific cognitive systems. While "pure" computational models are suitable for answering very general questions concerning the "power of AI" (questions (i),(ii) and (iii)), answering a more specific question (iv) and (v) will need a fairly evolved model of an embodied cognitive agent with a specific internal structure. Question (vi) will be answered with the help of answers (iv) and (v) and of yet another unconventional model of general computations.

What is the "Computational Power" of Artificial cognitive systems? In answering this question we are only allowed to exploit a minimal set of properties of cognitive systems on which majority of us agree. Minimality in this case means that removing any property from our list will result into a systems which could no longer be considered to be a typical cognitive system. It is generally agreed that the minimal set of such properties is: *interactivity*, enabling repeated communication of a system with its environment, to reflect environment's changes, to get the feedback, etc.; *evolution*, i.e., a development of a systems over its generations, and, last but not least, a potential *unboundedness over time* allowing an open-ended development of a cognitive system.

Note that classical Turing machines [1] which since Turing times have often been considered as "the computational model of mind" cannot model any fully fledged cognitive system – simply because such machines do not possess the above mentioned three properties. Hence their computational abilities and limitations cannot be considered to hold for cognitive systems.

Having in mind the above mentioned three properties of cognitive systems, in [2], [3] a very simple computational system – called *non-uniform evolving automaton* has been designed capturing precisely those properties.

Formally, a non-uniform evolving automaton is presented by an infinite sequence of finite–state transducers (FSTs). An FST is a finite-state automaton (FSA) working in a different input/output mode. Like any FSA, it is driven by its finite state control, but it reads a potentially infinite stream of inputs and translates it into an infinite stream of outputs. A non-uniform evolving automaton computes as follows: the computation starts in the first transducer which continues its processing of the input stream until it receives a so-called *switching signal*. If this is the case the input stream is "switched" over to the next automaton in the sequence. In general, a non-uniform evolving automaton is an infinite object. However, at each time a single transducer having a finite description is active. Switching among the transducers models the evolution of the system. The transducers in the sequence can be chosen in an arbitrary manner, with no classically computable relation among them. Thus, there might be no algorithm for generating the individual automata given their index in the sequence. This is why the evolution of the system is called non-uniform. In order to better model the "real" cognitive systems we may require that a specified subset of states of a given transducer is also preserved in the transducer in the sequence. In the language of finite transducers this models the persistence of data over generations of transducers. The switching signals are issued according to the so-called *switching schedule* that again can be a classically non-computable function. It comes as no surprise that a non-uniform evolving automaton, possessing non-computational elements, is a more powerful computational device than a classical Turing machine. For more details and the proof of the last claim, cf. [4]. Thus, the answer to the first question is that *interactive, non-uniformly evolving, and potentially time-unbounded cognitive systems (be it real or artificial ones) posses a super-Turing computing power: they cannot be modelled by classical Turing machines.*

Unfortunately, the super-Turing computing power of non-uniform evolutionary cognitive systems cannot be harnessed for practical purposes – it is only needed to precisely capture their computational potential, where the elements of uncomputability enter computing via unpredictable evolution of the underlying hardware and software.

Are There "Levels" of Intelligence? For answering this question we shall again consider the computational power of cognitive systems modelled by a

non-uniform interactive automaton. Namely, for such automata one can prove that *there exist infinite proper hierarchies of computational problems that can be solved on some level of the hierarchy but not on any of the lower levels* (cf. [5]).

The interpretation of the last results within the theory of cognitive systems is the following one. There exist infinite complexity hierarchies of computations of cognitive systems dependent on the amount of non-computable information injected into such computations via the design of the members of the respective evolving automaton. The bigger this amount, the more non-uniform "behaviors" (translations) can be realized. Among the levels of those hierarchies there are many levels corresponding formally (and approximately) to the level of human intelligence (the so–called Singularity level – cf. [6]) and also infinitely more levels surpassing it in various ways. The complexity classes defining individual levels in these hierarchies are partially ordered by the containment relation.

What Is the Position of Human Intelligence w.r.t. the "Levels" of Intelligence? There is increased theoretical evidence that the computational power of human intelligence (aided by computers or not) is upper bounded by the Σ_2 level of the Arithmetical Hierarchy.[1] This level contains computations which are recursive in the halting problem of the classical Turing machines. For instance, Penrose [8] argues that human mind might be able to decide predicates of form $\exists_x \forall_y P(x, y)$, i.e., the Σ_2 level. The computations within this class can answer the following question related to the halting of the arbitrary (classical) Turing machines for any input: (*"Does there* exist *a Turing machine which* for all *Turing machines and* for all *inputs decides whether they halt?"*). Similar conclusions have been reached during the last few decades by a number of logicians, philosophers and computer scientists looking at the computations as potentially unbounded processes (cf. [9]).

A more detailed structural insight into the nature of computations in the Σ_2 level of the Arithmetical Hierarchy offers a recent model of van Leeuwen and Wiedermann [9] – so called *red-green Turing machines*. This model characterizes the second level of Arithmetical Hierarchy in terms of a machine model.

[1] Arithmetical Hierarchy is the hierarchy of classically unsolvable problems of increasing computational difficulty. The respective problems are defined with the help of certain sets based on the complexity of quantified logic formulas that define them (cf. [7]).

A red-green Turing machine is formally almost identical to the classical model of Turing machines. The only difference is that in red-green Turing machines the set of states is decomposed into two disjoint subsets: the set of green states, and the set of red states, respectively. There are no halting states. A computation of a red-green Turing machine proceeds as in the classical case, changing between green and red states in accordance with the transition function. The moment of state color changing is called *mind change*. A formal language is said to be recognized if and only if on the inputs from that language the machine computations "stabilize" in green states, i.e., from a certain time on, the machine keeps entering only green states.

The model captures informal ideas of how human mind alternates between two states (accept and reject) when looking for a solution of a difficult decision problem.

Thesis 1 *The computational power of cognitive systems corresponding to human-level intelligence is upper-bounded by the class Σ_2 of the Arithmetical Hierarchy.*

Note that the previous thesis does not claim that the cognitive systems can solve all problems from Σ_2. Nevertheless, the example of the halting problem theorem shows that occasionally human mind can solve specific problems that in general belong to Σ_2 (for more details cf. [10]).

Is There a General Mechanism Behind the Human-Like Intelligent Systems? This is a very hard question, indeed. It can again be approached from the viewpoint of computations. If there were a different mechanism of intelligence than that we are aware today then there would be a notion of computation different from that we know about today. Note that we are speaking about computations, not about the underlying mechanisms. For all we know about computations today, there are many kinds of computations (deterministic, non-deterministic, randomized, quantum) each of which is characterized by a class of computationally equivalent mechanisms. We believe that this is also the case of cognitive systems which are but specialized non-uniform evolutionary computational systems supplied by information delivered, thanks to their own sensors and effectors, from their environment. (It is their environment that injects the non-uniform information into such systems, and their non-uniform development is further supported by Darwinian evolution.) Thus, one may characterize the mechanism of intelligent systems as any computational mechanism generating the class of computations (resulting further into behaviors) that those systems are capable to produce or utilize. For instance,

for such a purpose non-uniform evolving automata will do. However, we are interested in a more refined, more structural algorithmic view of cognitive systems possessing high–level mental qualities, such as learning, imitation, language acquisition, understanding, thinking, and consciousness. What are the main parts of such systems, what is their "architecture", what are the algorithmic principles behind their operation?

The answer is offered by the high level computational models of cognitive agents aiming at capturing higher–level human–like mental abilities. Among them, the most advanced modes seems to be the model named HUGO (cf. [10]) (cf. Fig. 1) which is conformed with the recent state of research in the domain of embodied cognitive systems.

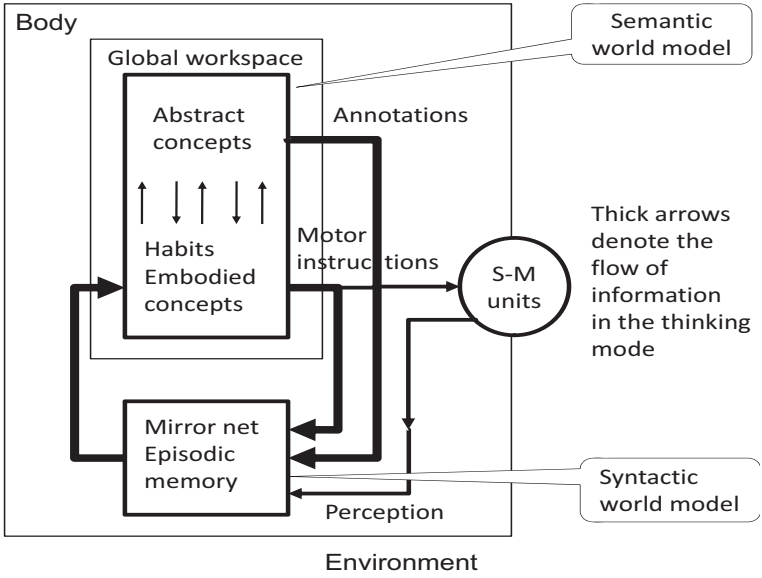

Fig. 1. The structure of a humanoid cognitive agent (HUGO)

The notable part of the scheme in Fig.1 is the body represented by the sensory–motor units. These units are governed by the control unit consisting of two main parts called *syntactic* and *semantic world model,* respectively. These two world models are realized with the help of neural nets and are automatically built during the agent's interaction with its environment. The syntactic

world model builds and stores the "database" of frequently occurring *multimodal units*, i.e., of tuples of sensory information and motor instructions that "fit together", make sense under circumstances corresponding to the given perception and proprioception. This database can be seen as a vocabulary of atomic units of behavior that have turned out to be good in the past. The semantic world model connects multimodal units into a semantic net that captures often followed sequences of activations (usages) of individual multimodal units. In the series of papers [11], [12], and [10] algorithmic mechanisms are described leading to the algorithmic emergence of higher mental abilities, such as imitation, language development and acquisition, understanding, thinking, and a kind of computational consciousness.

HUGO is not a universal high-level scheme of a humanoid cognitive system in the sense that it could simulate any other such system (like a universal Turing machine can simulate any other machine). This is because HUGO involves embodiment and (thus) morphology (albeit indirectly, via properties of sensorimotor units), and such aspects make the respective cognitive systems unique (for instance, one cannot simulate birds on fish).

Obviously, there might exist other "schemes" of humanoid cognitive agents, but the "validity" of the one we have presented is supported by the fact that, unlike the other schemes, it offers plausible explanation of a full range of mental faculties. Any other scheme with the same range would necessarily be equivalent to HUGO.

Can "Fully-Fledged" Body-Less Intelligence Exist? With the only exception of HUGO the previous models of cognitive systems were general, "disembodied" computational models capturing certain aspects of cognitive systems which we showed were enough to support the answers to our questions. Nevertheless, HUGO has been the only computational model for which we have been able to design algorithmic mechanisms arguably supporting the development of intelligence. For this to happen it was crucial that we have considered a complete cognitive agent inclusively its body represented by its sensorimotor units. The body has been an instrumental part of our agent allowing him not only to interactively learn his environment (to make himself situated in it) and thus, to build his internal structures (most notably the syntactic and semantic world model and episodic memories) on the top of which higher mental abilities have arisen so to speak "automatically" (cf. [12]). Agent's understanding of its own actions and perception has been grounded in the multimodal concepts formed by his sensorimotor units. From this viewpoint, the remaining models, lacking the body, could at best be seen as seriously

crippled models of cognitive agents. Could such purely computational, bodyless models retain the cognitive abilities of the embodied models of cognitive systems? It seems that contrary to popular beliefs that embodiment is condition *sine qua non* for intelligent agents, this belief is only partially warranted. Namely, according to the "theory" behind the HUGO model, embodiment is necessary in order intelligence to develop. However, once the necessary structures (and again, most notably the internal world models and the episodic memories) are developed, the agent (e.g., HUGO) can be *de-embodied*. That is, all its sensory-motor units can be removed from it, except those serving for communication (speaking/hearing or reading/writing). The resulting agent will work in the "thinking mode" using the cycle denoted by thick arrows in Fig. 1, being not able to develop any new skills and concepts related to sensorimotor activities. The de-embodied agent will "live" in a simulated, virtual world provided by his internal world models. His situation will thus remind the circumstance described in the philosophical thought experiment "brain in the vat" (cf. [13], [14]).

Can There Be a Sentient Cloud of Gas? Written by by astrophysicist Sir Fred Hoyle the nowadays cult science fiction novel "The Black Cloud" [15] appeared in 1957. When observed from the Earth, this cloud appeared as an intergalactic gas cloud threatening to block the sunshine. After a dramatic attempt to destroy the cloud by a nuclear bomb the scientists came to a conclusion that the cloud possessed a specific form of intelligence. In an act of a pure hopelessness, they tried to communicate with it and, to their great surprise, they discovered a form of life, a super–organism obeying intelligence surpassing many times that of humans. In return, the cloud is surprised to find intelligent life-forms on a solid planet.

By the way, extra–terrestrial sentient oceans, planets, and suns occur quite often in numerous sci–fi novels.

How plausible is the existence of such sentient super–organisms? To answer this question we will invoke another result related to non-standard machine models of computations – so-called *amorphous computing systems*. From a computational viewpoint, amorphous computing systems differ from the classical ones almost in every aspect. They consist of a set of similar, tiny, independent, anonymous and self-powered processors or robots that can communicate wirelessly to a limited distance. The processors are simplified down to the absolute necessaries in order to enable their massive production. The amorphous systems appear in many variants, also with nano-sized processors. Their processors can be randomly placed in a closed area or volume and form

an ad-hoc network; in some applications they can move, either actively, or passively (e.g., in a bloodstream). Depending on their environment, they can communicate either via radio, via signal molecules, or optically, or via whatever wireless communication means. The investigation of such systems has been initiated by the present author by the beginning of this century (for an overview, cf. [16]). Amorphous computing systems appear in many forms and the simplest ones can consist of processors which are, in fact, simple constant depth circuits. Genetically engineered bacteria can also be turned into an amorphous computing system [17]. The main result that holds for such models is that all of them they possess universal computing power. This means that they can simulate whatever computation of a classical Turing machine. For the simplest amorphous computing systems such a simulation is unbelievably cumbersome, because the underlying amorphous computing system can compute but with the unary numbers. This will cause an exponential slow-down w.r.t. the original computation.

Now we are in a position to formulate the answer to the question of this subsection. The "cloud" can be seen as a specific amorphous computing system. According to what has been said previously, such a system can simulate the computational part of, e.g., HUGO that was mentioned in the previous subsection. The whole super–organism will not be completely body–less, since its processors have locomotion and communication means, and possibly other sensors and actuators. According to what we know the cloud will be able, over the entire existence of the Universe, develop a form of intelligence that will be appropriate to the environment in which it lives. The "slowness" of its thinking does not matter, taking into account travel time needed to investigate the potentially unbounded space. Undoubtedly, Darwinian evolution will also apply to this case. Interestingly, recently physicists have discovered inorganic dust with life-like qualities [18].

And could such a cloud be many times more intelligent than people? This is hard to say because its intelligence will be of a different nature than ours. But the principles of evolution and operation of its intelligence will be the same as those of us. Computational arguments can again be invoked showing that even an amorphous computing system of galactic size will not be able to solve problems beyond the Σ_2 class of the Arithmetic Hierarchy (cf. [10]).

3 Conclusions

We have seen that using the non-standard machine models of the contemporary theory of computations and the current ideas on the working of non-trivial cognitive systems we are able to answer the questions that until recently

have been the domain of sci–fi or of philosophy, at best. On one hand, the answers deny the ideas of some sci–fi writers or of some prodigies of science (cf. [6]) concerning the existence of super–intelligence. On the other hand, they also support futuristic ideas concerning the development of alien intelligence in alien environments using alien forms of life. It is encouraging to see how seemingly unrelated theories of non-standard models of computations and theory of cognitive systems go hand in hand in our quest for unraveling the secrets of intelligence.

Acknowledgments. This work was partially supported by RVO 67985807 and the GA ČR grant No. P202/10/1333

References

1. Turing, A. M.: On computable numbers, with an application to the Entscheidungsproblem, Proc. London Math. Soc. 42(2), 230–265 (1936); A correction, ibid., 43, 544–546, (1937)
2. van Leeuwen, J., Wiedermann, J: The Turing machine paradigm in contemporary computing. In: Enquist, B., Schmidt, W. (eds.) Mathematics unlimited - 2001 and beyond, pp. 1139–1155. Springer (2001)
3. van Leeuwen, J., Wiedermann, J.: Beyond the Turing limit: evolving interactive systems. In: Proc. SOFSEM'01. LNCS, vol. 2234, pp. 90–109. Springer, Berlin (2001)
4. Wiedermann, J, van Leeuwen, J.: How We Think of Computing Today. (Invited Talk) In: Proc. CiE 2008. LNCS, vol. 5028, pp. 579–593. Springer, Berlin (2008)
5. Verbaan, P.R.A., van Leeuwen, J., Wiedermann, J.: Complexity of Evolving Interactive Systems. In: J. Karhumäki et al. (eds.) Theory Is Forever – Essays Dedicated to Arto Salomaa on the Occasion of His 70th Birthday. LNCS, vol. 3113, pp. 268–281. Springer, Berlin (2004)
6. Kurzweil, R.: The Singularity is Near. Viking Books (2005)
7. Cooper, S. B.: Computability Theory. Chapman&Hall/CRC (2004)
8. Penrose, R.: Shadows of the Mind: A Search for the Missing Science of Consciousness. Oxford University Press, Oxford (1994)
9. van Leeuwen, J., Wiedermann, J.: Computation as Unbounded Process. Theoretical Computer Science (2012)
10. Wiedermann, J.: A Computability Argument Against Superintelligence. Cognitive Computation 4(3), 236–245 (2012)
11. Wiedermann, J.: HUGO: A Cognitive Architecture with an Incorporated World Model. In:Proc. of the European Conference on Complex Systems ECCS'06. Said Business School, Oxford University (2006)

12. Wiedermann, J.: A high level model of a conscious embodied agent. In: Proc. of the 8th IEEE International Conference on Cognitive Informatics, pp. 448–456 (2009); expanded version appeared in: International Journal of Software Science and Computational Intelligence (IJSSCI) 2(3), 62–78 (2010)
13. Wiedermann, J.: Towards Computational Models of Artificial Cognitive Systems that Can, in Principle, Pass the Turing Test. (Invited talk) In: Proc. SOFSEM 2012. LNCS, vol. 7147, pp. 44–63. Springer, Berlin (2012)
14. Wiedermann, J.: On the Road to Thinking Machines: Insights and Ideas. (Invited session talk) In: Proc. CiE 2012. LNCS, vol. 7318, pp. 733–744. Springer (2012)
15. Hoyle, F.: The Black Cloud. Penguin Books (1957)
16. Wiedermann, J.: The Many Forms of Amorphous Computational Systems. In: Zenil, H. (ed.) A Computable Universe: Understanding Computation & Exploring Nature As Computation. World Scientific Publishing Company (2012) to appear
17. Wiedermann, J.: Nanomachine Computing by Quorum Sensing. In: Kelemen, J., Kelemenová, A. (eds.) Computation, Cooperation, and Life. Essays Dedicated to Gheorghe Păun on the Occassion of His 60th Birthday. LNCS, vol. 6610, pp. 203–215. Springer (2011)
18. Tsytovich, V. N. et al.: From plasma crystals and helical structures towards inorganic living matter. New J. Phys. 9(263) (2007)

Is Evolution a Turing Machine?

Vít Bartoš

Technical University of Liberec, Liberec, Czech Republic
vit.bartos@tul.cz

Abstract. This article deals with the basic question of the design principles of biological entities and artificial ones expressed by Gerald Edelman's question "Is evolution a Turing machine"? There is a general belief asserting that the main difference between evolutionary computation and Turing model lies in the fact that biological entities become infinitely diverse (analog) and fundamentally indeterminate states. I am of the opinion that this difference is not the issue. Because the nature has on its elementary level quantum structure which is therefore basically digital. Differentiation between evolution and human-formed machine lies in the physical structure of biological entities linked to the scaling of all physical levels. This architecture works as multi-domain value system whose most basic function is the categorization of events entering the field of interaction of the organism. Human thinking as a product of evolution is a prime example of this process. Processes of Turing machine simulate only a certain aspect of thinking and are not able to implement many others. Evolution, therefore, is not Turing machine.

Keywords: Evolution, Turing machine, Leibniz, physical structure, hierarchy, logical structure, value system, categorization, analog, digital, quantum, scale structuring, engineering approach, biological approach

1 Engineering and Biological Models

François Jacob claimed that in terms of constructional structure of things biological evolution[1] should be understood as a work of a handyman while

[1] By the biological evolution we generally mean the process of these essential stages: there is a common ancestor; there is a variation in genes, genotypes, phenotypes; there is a multidimensional selection basically on the level of phenotypes (but as a consequene there is a selection on other levels); finally there is a heredity of favoring features.

the artificial objects of human culture should be envisaged as the work of an engineer. This metaphor tells us simply that the engineer works with the precisely defined entities while evolution does not know anything like that and builds on what is at hand and also spontaneously.

Engineering or cybernetic model of the human mind is historically linked with the notion that the essence of human thinking is logical operations with the given symbols. In modern terminology this position is called cognitivism:

> The central intuition behind cognitivism is that intelligence – human intelligence included – so resembles computation in its essentials characteristics that cognition can actually be defined as computations of symbolic representations. [1](p. 40)

The cognitivist approach implies an interesting consequence. Anything that performs logical operations with symbols should be understood as the rudimentary beginning of intelligence. Human intelligence is not substantially different from any machine performing logical operations with symbols; it's just a question of computing power, memory and information processing time. When we ascribe the fact that the logical operators can be implemented in virtually any substrate material the conclusion that the mind (intelligence) is not significantly dependent on biological structures could be done. This laid the foundations of functionalist theory of multiple realizations (substrate variation) of the function or the logical structure. Turing machine (a combination of finite state automata, and infinite tape) thus represents an ideal model to which any physical system operating in a limited variety of operations and discrete states can be reduced. Therefore, there is the only one type of universal computing machine.

Gerald Edelman puts a provocative question that defines sharp distinction between these two models: "Do you think that evolution is a Turing machine?" According Edelman's vision – neuronal Darwinism – the human thinking process is very similar to natural selection – there are not instructions here; there are not clear and discrete states, which are the finite number as in the case of digital machines. States and operations of the real biological system (the brain) cannot be sharply defined. They are in fact blurred (fuzzy) because they are necessarily contextual:

> We cannot individuate concepts and beliefs without reference to the environment. The brain and the nervous system cannot be considered in isolation from states of the world and social interactions. But such states, both environmental and social, are indeterminate and

> open-ended. They cannot be simply identified by any software description. [2](p. 224)

In fact cognitive processes are fundamentally based on perpetual natural selection among groups of neurons (groups of representations) which are temporarily set up in response to a current problem and which are constantly transforming. An important part of this global process is also creating a reciprocal feedback loops (reentry) that integrate functionally separate areas of the brain and generally coordinate the interaction between value systems.

With the above mentioned there is closely related issue of continuity and discreteness conditions in biological structures:

> Now we begin to see why digital computers are a false analogue to the brain. . . . The tape read by a Turing machine is marked unambiguously with symbols chosen from a finite set in contrast, the sensory signals available to nervous systems are truly analogue in nature and therefore are neither unambiguous nor finite in number. [2](p. 224)

Edelman claims explicitly that there is almost "ontological" difference between artificial and biological entities. Artificial objects operate on the atomic discrete states (characters on the tape Turing machines) whereas biological entities operate on a range of values of the continuum (expressible in real numbers). In this case there is obvious consensus between Turing and Edelman because Turing claims as well:

> The nervous system is certainly not a discrete-state machine. A small error in the information about the size of a nervous impulse impinging on a neuron may make a large difference to the size of the outgoing impulse. It may be argued that, this being so, one cannot expect to be able to mimic the behavior of the nervous system with a discrete state system. [4](p. 456)

From the ontological point of view the problem of biological and artificial systems is extremely important and its examination will probably explain a number of uncertainties which we have described above.

With your permission, I switch right now for a while on the level of basic metaphysical problems. It may look at first glance like a superfluous thing, but I suppose that the basic metaphysical (ontological-system) intuitions play in our human thinking and science quite a substantial role.

I would like to submit here now one problem and one resolution that Gottfried Wilhelm Leibniz formulated in the early 18th century. The first

problem concerns two most fundamental questions that people ask, while one of them is related to our problem. I shall try to answer it very shortly, because the answer will form the basis of our consideration of the relationship of analog and digital.

Further, let us recall Leibniz's distinction between artificial creations and divine creations (natural creations). This heuristic resolution supports the generality of our further scaling theory of the structuring and interconnection products of a process of biological evolution. Let's start with these major problems. Leibniz explicitly formulates them and I am convinced that the value of these questions can hardly be overestimated (italics added by the author of this paper):

> There are two famous labyrinths where our reason very often goes astray: one concerns the great question of the Free and the Necessary, above all in the production and the origin of Evil; the *other consists in the discussion of continuity and of the indivisibles which appear to be the elements thereof, and where the consideration of the infinite must enter in.* The first perplexes almost all the human race, the other exercises philosophers only. [3](p. 54)

We will now be interested in the second labyrinth, concerning the relationship between continuum and discretion, which are opposite possible properties of basic ontological structures, such as time, space and matter, or in modern times the information (meaningful, identifiable difference). I defend the view that the essence of physical reality are discrete entities. There are the empirical and hypothetical reasons for which I reckon discovery and prediction of modern experimental and theoretical (quantum) physics.

But there are, in my opinion, the reasons a priori. Perfect continuity (cognitively modeled as a continuous interval, Euclidean plane or Cartesian homogeneous space and formally described by the concept of real numbers) entity excludes difference between things. Exclusion of difference (information) makes it impossible to application of the principle of sufficient reason (in Leibnizian terms told). And if there is no sufficient reason, there can be anything happening, or vice versa anything cannot be happening at all. Leibnizian units of reality, called "monads" are therefore individualized, because they prevent from the perfect homogenity – or in modern terms, from the absence of information.

The conclusion is that, strictly speaking, only discrete entities can exist. All existing systems with a finite number of discrete elements then behave digitally and can be understood as finite automata. This universal rule, of

course, implies that the biological systems are finite automata as well. This conclusion comports with the engineering approach and is in stark contrast to the biological concept. Refusal to understand biological entities like machines (automata) is deeply embedded in our imagination and has its intellectual and emotional context that is humanly understandable. I would only say that the identification of biological entities with machines actually does not diminish the value of the natural world. In fact it depends on the actual physical architecture and scaling structuring and consistency in other words, on the complexity of these machines. That and this reflects the 64th Leibnizs Monadologie thesis, where a distinction is made between two types of machines – machines created by humans and machines created by God – in today's terminology, by nature or evolution (again, the italicized parts were emphasized by the author):

> Thus every organized body of a living thing is a kind of divine machine or natural automaton. It *infinitely surpasses any artificial automaton*, because *a man-made machine isn't a machine in every one of its parts.* For example, a cog on a brass wheel has parts or fragments which to us are no longer anything artificial, and bear no signs of their relation to the intended use of the wheel, signs that would mark them out as parts of a machine. But *Nature's machines – living bodies, that is – are machines even in their smallest parts, right down to infinity.* That is what makes the difference between nature and artifice, that is, between divine artifice and our artifice.[2]

Now, when we abstract from the historically contingent conceptual constructs of "divine machine" and from the assumption of infinite structuring systems (impossible in terms of thermodynamics and control), we get constructive hypothesis about the difference between artificial and natural automata. The Leibniz's hypothesis simply says that the natural (living) entities unlike artificially constructed entities are machines even in their parts, and so it works across physical systems of all space-time levels (in modern interpretation).

Subsequent considerations are essentially based on just those originally Leibnizian concepts – they are just upgraded explications of these ideas. Deduced consequences, largely reconciling biological and engineering approach – we see as proving the genius of Leibniz's formulation.

[2] Available on: `http://www.earlymoderntexts.com/f\_leibnitz.html`

2 Analog or Digital

As we have seen above there are shared intuitions about the diversity of nature of states and transitions logic between the states in biological and artificial entities. Turing machine tape with its discrete coded and clearly defined states is at first glance something different than comprehensively multi-domain and fuzzy states such as the nervous system. Algorithms are absolutely something different than natural selection.

When thinking about the issue we will have to come down to a completely elementary level of physical reality – in microcosm as its entities are at the base of all existing things. In simple terms: quantum world is close to the digital world. It appears that the mass and energy in the last instance exist only in discrete portions (Planck's domain). According to some extravagant interpretations even space-time and motion are quantized – i.e. discretized. In this case our problem would be easily solvable – fuzziness conditions in biological domains are given of our own – needless to say principal – ignorance, our inability to distinguish reality of the finest domains and their overlapping or inclusion in the hierarchy of complex physical systems. Fuzziness is only an illusion in fact or in terms of "God's eye view", every system is perfectly defined through conditions of "status" atoms-quantum physics grid. Everything that exists could then be seen as a "discrete-state system", i.e. a system that resembles a Turing machine.

The first thing we should solve is question of what it means to change the state of the system or switch from one system state to a different one? The change of something called the state of the system must be a relevant change. The word "relevant" refers to any significant change in internal or external relations (symmetry or asymmetry) part of that entity. It's hard to believe that in a true "continuum-state machine" (analog machine) meaningful state transformation occurs in just one single position within a continuous interval of transition between states. If it be true then the structural change would be infinitely sensitive to the correct input which is critically unlikely. The opposite extreme would be a statement that the structural change in the system can be considered as any mechanical change in the position of any parts of the system. Then by the slightest movement of any of its part the system should go through endless systemic transformations which is absurd as well.

Provided the strictly analog process, system in transition, should require the infinitely precise identifiers of change which is impossible. This is confirmed by Daniel Hillis:

> Although we have an infinite number of possible values of the signal, only a finite number of values are of a meaningful difference – therefore represents information. Doubling the number of meaningful differences in the analog computer would do everything twice as accurate ... [5] (p. 66)

From what has been said the following implies: Strictly analog process is a fiction. Relevant information causing change in the system state must occur at specific intervals of values factually relating to the scale structuring and complexity of an entity. If the relevant information necessary for the state change can occur in the finite intervals of values only then this is a digital process. Structural change in the system – the transition from one state to another – is necessarily discrete matter. If it were not so there would be the system either infinitely sensitive to incoming signal (waiting for one single value on the interval of real numbers) or vice versa unable of distinguishing one value from the other and completely insensitive to the intensity of the signal – because of absence of sufficient reason for a choice. Only a discrete portion of the signals and discrete states of systems represent a meaningful entity capable of interacting within a limited behavior variety.

There is not any fundamental distinction between the Turing machine and the evolution – with respect to discrete or continuity information structure of entity. In fact the notion of information necessitates discrete states.

3 Hypothesis of Scale Structuring and Interdependence

Perhaps we should ask ourselves why the states of biological systems seem us actually ever analog and not digital. When both Edelman and Turing argue that the nervous system and brain are sensitive to small changes in signals and environmental context then it looks like a very rational justification for analog communication structure. We were able however to show that provided the quantum structure of the world and the concept of meaningful difference (information for interacting system) given there exist de facto discrete (digital) systems only. The phenomenon of states fuzziness especially for biological entities is due, in my opinion, to what I would call scaling linkages of physico-biological domains. I mean the scale linkages to be a simple fact that biological entities in themselves contain a hierarchical cascade of physical entities from elementary particles, molecular and macromolecular structures, cells, organs and organisms to ecosystems. The interdependence of these domains is very complex and reciprocal. This means that the state of the biological entity is

in fact a complex scaling – domains of different size, complexity and duration are overlapping. This overlap – which is only partially empirically detectable – is the cause of putative blurriness of states of biological entities.

Personally, I believe that the human mind as a biological phenomenon is a prime example of this process. The assumption of global interdependence scaling biological entities derives significant results! Let us compare them with the engineering approach: Engineering approach bases its strategy on the separation of the logical structure and physical structure of the entity which is the basis of functionalist theory of multiple implementation of the object (function). Simply said it does not matter what are logic gates and a substance that is to go through them. Implementation of Boolean logic is the substrate (material) neutral. The second problem of the engineering approach lays in abstracting from the fine consistency of hierarchical architecture of natural objects. In practice the construction of artificial entities mimicking biological entities abstracted from a certain level of organization – e.g. artificial neural networks is abstracted from a lower level of real processes taking place inside the cell of real neuron (this may miss additional computing capacity of a biological system). The result of this type of approach is the concept of intelligence (mind), which is not delimited by the space-time frame (no matter how slowly can logical operation proceed on no matter how large entity) and completely abstracted from the real hierarchical composition (complexity) of physico-biological entities.

Biologists are clearly against this concept. The real biological system and therefore real thinking clearly matters on the spatio-temporal and compositional characteristics of entities. Logical architecture of biological systems is not separable from their physical level. This means that what we call "logical operations" and what we model as a physical structure of the gates through which any substance flows is abstraction. The absurdity of this abstraction quickly realizes when we consider well what it means to abstract from the composition and spatio-temporal properties of entities. In the terms of the traditional philosophy it should mean abstraction from the primary qualities of an object which is the same as to say that an object A with certain essential characteristics is the same object as object B which does not have these essential qualities. This is obviously absurd assertion. In the terms of physics this should mean abstracting from thermodynamic determination of physical systems just like from the obvious (space-time) scale dependent position and function of each specific physical entity in relationships with other physical entities. Finally, in the area of semantics this should mean abstracting from the fact that the meanings of terms are introduced in limited field of significance – meanings are necessarily anthropometric. Excessive inflation of this

field leads to the complete degradation of the original meaning. For example, if you intend to adjudge the term "thinking" to objects of completely different physical structure than the intelligent mammals, the question is whether has the term "thinking" still any differentiating sense in such an extremely liberal-established language game.

Generally expressed: an engineering approach commits cognitive misconduct – something what Alfred North Whitehead called "the Fallacy of Misplaced Concreteness". This means nothing else than that we as human beings are prone own abstractions considered as an adequate expression of reality.

4 Biological Architecture–Value Systems

I consider that what we call "thinking", as clearly biological phenomenon. Reducing the thinking to mathematical reasoning ability and purely verbal response – i.e. to the symbolic activity, as Turing did, is probably inadequate. Biological machines must firstly follow evolutionary logic that is unconsciously and independently of the level of biological control domain imperative: "Survive, preserve yourself, replicate!" In addition to this, the hard fact that our world is an irreversible process where the slightest change (butterfly effect) can have fatal consequences for a particular organism in real-time we find the fact that biological organisms must be in the first place machines able to flexible response and reception in a real time in a wide range of physical effects. For better understanding to the logic of biological entities we have to admit one more assumption – in our type of universe there are objects arranged hierarchically with a certain asymmetry in the interaction between domains. I call them "asymmetrical relations". The principle is simple: the elementary level strongly determines the emergent ones and not vice versa. As an example consider the question of the necessary conditions for the existence of complex entities (e.g. life). Positive stability of certain elementary particles and the structure of molecular complexes is a necessary condition (besides numerous others) for the existence of living beings on the suitable planet. But not vice versa – elementary particles and molecules will exist independently of the existence of life. Therein lies the asymmetry. This asymmetry is also valid for other scales of physical systems and of course on the level of complex biological systems.

If this principle seems to be inconclusive or incomprehensible to you, think of the problem as an illustration of the principles of Lamarckism – in particular the principle of inheritance of acquired characteristics of organisms (their transmission to offsprings). This thesis is not only empirically proven as incorrect, but also represents a logical and systematic problem, as shown e.g. by

Gregory Bateson. If the experience of the individual organism in a changing environment could transmit directly to offsprings, it is necessary to admit a number of absurdities. Here are some examples:

- Experience is in an individual organism during its life often contradictory – it means that it is then possible to have a completely contradictory adaptation as acquired properties?
- Adaptation variety could be potentially endless – just like individual differences within a species that exist in a variable and irreversible environment.
- What ever the term "species" means, if each individual can produce such somatically very different offspring? How is ensured the compatibility of the mating organisms in the process of the sexual reproduction?
- With what frequency are various adaptations changed – how many members must have an inductive series of experience leading to a new adaptation? What system assesses the inductive experience as sufficient to change the properties of an organism?
- How is provided the compatibility of acquired property with other properties? Etc.
- How are the organism regulatory circuits functioning? Homeostatic balance (range of values of variables) is possible only if there is determinative metasystem (privileged modular structure). Metasystem however implies asymmetry links!

The essence of Lamarckism lays in assumptions that basically everything is possible, or at least it is not obvious what the fundamental limitations of the organism to acquire new properties are. If we were able to consider Lamarckism vision to reductio ad absurdum, there would be no restriction on the transformation of organisms, except the external constraints. But Lamarckism principle can be applied (recursively) on these limitations and then after a generalization we get the intolerable conclusion that anything can be transformed in any way. Lack of system privileged relatively invariant structure, capable to restrict variety in behavior of emergent layers, leads to the above mentioned consequences. Where there is no hierarchy in the arrangement of the system, there are fails in order organization of the relevant processes. Terms such as "greater or lesser importance" for such a system make no sense. But this is absolutely not any of our experience with the systems of nature. Absence of hierarchically organized domains of physical reality would cause the collapse of the principle of sufficient reason – the unthinkable chaos, or, conversely, the inability of the transition from present state to the following one. These

are Leibnizian conclusions that strike me as resilient, although I admit that I could be mistaken.

Therefore the principle of asymmetrical relation that expresses the system principle of physical reality should be accepted, despite the fact that the metaphor of the hierarchical structure of reality, which implies a binding principle of asymmetry, seems in many respects to be outdated or naive.

After all a simple conclusion is following: biological systems (including human thinking) are designed by natural selection as categorical systems, or, if you please, the value architecture. This means that in the asymmetrically coupled and hierarchically organized universe each event through organism perceived has a certain degree of relevance. Organisms had to learn to categorize and sort the events of the physical world according to the degree of importance due to their own existence. Let's call this process "evaluation events" and cognitive architecture body corresponding "value systems" (Edelman's term).

Results of an evolutionary proces – an evolutionary computation – are therefore the value systems of the organism whose task is multidomain assessment of the situation (categorization) in which the entity is located, and then decide what to do for self-preservation of the organism first.

I believe that the essence of thinking (to what extent is the thinking inherently biological phenomenon) is the assessment of events, categorization, which cannot be implemented on a Turing machine. Why? Because Turing machine is not any value system from the nature of its physical structure and we have agreed that physical constraints are important. The problem ultimately lies not in question whether states are discrete entities or analog dependent. Calculations on the value systems are discreet as well as on idealized Turing machine, but are parallel on many different space-time domains (from microstructures cells to mechanical parts of the body) and are scales linked.

Therefore evolution is not any Turing machine.

References

1. Varela F.J., Rosch, E., Thompson, E.: The Embodied Mind. MIT Press, Massachusetts (1991)
2. Edelman, G.: Bright Air, Brilliant Fire. BasicBooks, New York (1992)
3. Leibniz, G.W.: Theodicy. Open Court Publishing Company, Illinois (1985)
4. Turing, A.: Computing Machinery and Intelligence? In: Copeland, J. (ed.) The Essential Turing. Clarendon Press, Oxford (2004)
5. Hillis, D.: The Pattern on the Stone. Basic Books, New York (1999)

From Science Fiction to Social Reality

Jelena Guga

University of Arts in Belgrade, Serbia
hrast78@gmail.com

Abstract. The emerging technological developments across various scientific fields have brought about radical changes in the ways we perceive and define what it means to be human in today's highly technologically oriented society. Advancements in robotics, AI research, molecular biology, genetic engineering, nanotechnology, medicine, etc., are mostly still in an experimental phase but it is likely that they will become a part of our daily experience. However, human enhancement and emergence of autonomous artificial beings have long been a part of futures imagined in SF and cyberpunk. While focusing on the phenomenon of cyborg as a product of both social reality and fiction, this paper will attempt to offer a new perspective on selected SF and cyberpunk narratives by treating them not only as fictions but as theories of the future as well.

Keywords: cyborg, science fiction, cyberpunk, bodily enhancement, artificial organisms, holograms, memory

Throughout the history, with every technological breakthrough, innovation or revolution, people have always imagined possible futures that new technologies at hand might bring about. In our predictions and projections of hopes and fears onto the future, literature, art and film have not only had an important role in shaping the ways we imagine the future of humanity, but have also prepared us to adapt to and gradually accept the ideas of technologically mediated existence thus incorporating them into the lived reality we share today. Mary Shelley's Frankenstein, Frank L. Baum's Tinman, Edgar Allan Poe's General Winfield Scott whose body is composed of prostheses, Fritz Kahn's illustrations representing human body as industrial machinery, Fritz Lang's film *Metropolis*, and Charlie Chaplin's *Modern Times* are only but a few of numerous examples of technologically augmented or enhanced bodies representing the merging of biological and artificial, natural and monstrous, human and machine, that can be found in the history of literature, visual arts and film and can be considered precursors

of cyborgs as we imagine and define them today. The proliferation of various modern cyborg forms imagined through art, fiction and popular culture emerged in the second half of 20th century along with (and as the reflection upon) the development of telecommunication technologies, military industry, entertainment industry, computer science, cybernetics, robotics, cognitive science, genetics, space travel explorations, advancements in medicine, digital imaging, etc.

Different representations of organic and technological merger were annotated different names such as bionic systems, vital machines, teleoperators, biotelemetry, human augmentation or bionics [1], until the introduction of the term "cyborg" which in 1960 became and still remains the common denominator of these phenomena. The term was coined by Manfred E. Clynes and Nathan S. Kline in the article "Cyborgs and Space" [2], and was used by the two scientists to describe the advantages of self-regulatory human-machine system adjustable to different environments invasive for the human body, that could as such be used for space travel.

As a theoretical concept, cyborg was then defined in terms of his/her/its abilities to deliberately incorporate "exogenous components extending the self-regulatory control function of the organism in order to adapt it to new environments." [2] (p.31) In demonstrating the feasibility of this idea, they presented the first cyborg which was neither a monstrous product of science fiction nor a cybernetic enhanced human being, but a mouse with a Rose osmotic pump implanted under its skin, injecting chemicals into an organism at a controlled rate thus creating a self-regulating closed system. Clynes and Kline suggested that the application of a similar system on astronauts could solve space travel problems such as fluid intake and output, cardiovascular control, blood pressure, breathing, perceptual problems, hypothermia, etc. in an automatic and unconscious way, "leaving man free to explore, to create, to think, and to feel." [2] (p.31) Speaking of such a perfect astronaut, these two scientists actually identified a new form of biotechnological organism that has ever since strongly influenced the ways we imagine, construct and define the body in relation to technological development.

Apart from being used to describe a perfect astronaut, the meaning of the term cyborg was broadened and widely used in both science fiction and scientific research to mark various forms of biotechnological couplings. However, it was only after the publication of now famous "Cyborg Manifesto" by Donna Haraway [3] that the notion of cyborg was given serious attention to in academic and nonacademic intellectual circles. Haraway recognized the potential of polysemous implications of the term and used it as a rhetorical and

political strategy to deconstruct ambiguous definitions of the subject within the postmodern digital culture. With a remarkable clarity and a certain dose of irony, she managed to outline a new provocative posthuman figure and initiate new philosophical and (bio)political orientations as well as disciplines such as cyborgology or cyborg theory which became central concepts not only for the work of academics in the field of technological development, but also for political scientists, military historians, literary critics, artists, computer scientists, sociologists, medical doctors, psychologists, philosophers and many other cultural workers.

In other words, Haraway's manifesto represents a milestone which opened up a new perspective in theoretical thought on how technologies impact and redefine the notion of human. Apart from showing the importance of Haraway's manifesto for the ubiquitous use of the term cyborg, I do not intend to reinterpret the manifesto all over again, since it has already been done by many prominent thinkers in the field of cyberculture studies, feminist studies, new media theories, as well as in cyberfeminist and other new media art practices. However, I will extract and throughout this paper intertextually entertain a thought from the manifesto which states that "the boundary between science fiction and social reality is an optical illusion." [3]

Through various examples, I will thus attempt to show how cyborg, not only as Haraway's theoretical concept or myth but also as an imaginary construct of fiction, has become a part of our present reality. Moreover, the boundary between the present and the future is now collapsing as never before, for we now live in a time when certain futures of science fiction that include ubiquitous networking, humanoid robots, artificially grown tissues and body parts, prosthetic extensions of the body, implants, AI, genetic modifications, alterations and crossbreeding, are palpable and have already become or are in the process of becoming the scientific and social reality of our present.

In other words, due to the exponential technological development we are witnessing today, the future and the present are now overlapping and intersecting in so many ways and are interwoven on so many levels, that William Gibson, a cyberpunk writer who coined the term "cyberspace", has a point when saying that the future is already here – it's just not evenly distributed. Future simply isn't what it used to be because it has become a part of the perpetual and extended "now" that we live in, or as Gibson has explained it in his novel *Pattern Recognition*:

> Fully imagined cultural futures were the luxury of another day, one in which "now" was of some greater duration. For us, of course, things can change so abruptly, so violently, so profoundly, that futures like

> our grandparents' have insufficient "now" to stand on. We have no future because our present is too volatile. [4]

As the technologies develop and change at an ever greater pace imposing the future upon us, the notion of cyborg is changing accordingly. For example, the rapid changes in cyborg representations is explicitly shown through the *Terminator* film franchise where in a bit more than twenty years timeframe, cyborg has transformed from the masculine coded rough, indestructible, unstoppable, aggressive and potent body, to an uncanny amorphous liquid metal that can take on any form, to female who, in the opinion of Saddie Plant have always been cyborgs [5], and finally to a cyborg who does not question or doubt his human existence because his biological brain and heart were implanted into a newly grown and constructed body without him being conscious about it. Cyborg transformation is still an ongoing process and therefore a unified or conclusive definition of cyborg does not exist. So instead of an attempt to define it at this point, I suggest outlining one of its key characteristics crucial for this paper: Cyborg is simultaneously imaginary concept and practical, material development of possible couplings between human (or any other organism) and machine, i.e. biological and technological. Roughly identified, the notion of cyborg can stand for an artificial body (robotic / synthetic) usually impaired with and governed by an AI, technologically modified and enhanced biological bodily and mental human capacities, or the combination of the two.

On phenomenological and ontological level, cyborg as a hybrid requires new ways of interpretation and articulation since its very existence as a single biotechnological entity redefines what it means to be human in a technologically mediated society where Cartesian dualisms or other essentialist concepts alike are not applicable. It is only through anti-essentialist theories (postmodernism, culture and cyberculture studies, theory of new media and new media art, etc.) combined with and/or applied to the works of science fiction, bio and transgenic artistic practices as well as scientific research, that we can only begin to comprehend and better articulate the influence and effects of these new forms of subjectivities that bring about radical changes in contemporary human experience.

Science fiction and especially cyberpunk with its dystopian visions of very near, almost palpable future, has proven to be more agile in keeping up with the pace of technological development than production of academic theoretical frameworks dealing with the impact of these phenomena, and very often preceding them. For example, remaking films such as *Total Recall*, *Judge Dredd*, and *In Time*, as well as negotiating remakes of *Ghost in the Shell*, *RoboCop*, *Dune*, etc., all show that we are more and more likely to turn

to a vast array of cyberpunkish future scenarios in order to better understand or figure out and cope with the technological cacophony of our present. So, for the purposes of this paper, insights of such writers as William Gibson and Philip K. Dick along with some important issues raised in carefully selected SF films, will be synchronized with theoretical and philosophical texts and treated as a theoretical framework that has a potential of deconstructing the distinction between science and fiction.

When discussing the changes brought about by new technologies, what should be taken into consideration is a distinction between those technologies that we encounter and use in everyday life and those that are currently being developed behind the closed doors of various scientific research centers and institutions and may or may not become a part of our daily experience. However, none of the two categories of the existing technologies should be dismissed or overlooked because their very existence raises important moral, ethical and other issues that matter to our human existence. These two categories of technological development very often overlap, but the distinction needs to be made in order to better understand the changes already brought about by ubiquitous use of new technologies and the potential changes we may witness most probably within a lifetime. With a reference to SF/cyberpunk texts, I will first address some of the already widespread interfacing possibilities, and then turn to several human augmentation experiments that bring science fiction future into the reality of present.

Interactions we have through our screens on daily bases are slowly giving way to newly created interfaces such as gestural interfaces (Nintendo Wii and Xbox Kinect gaming consoles, "g-speak" interface created for the purposes of film *Minority Report*, portable gestural interface "SixthSense" created by MIT's researcher Pranav Mistry, etc.), holographic projections (virtual assistants at Luton airport, projections of celebrities usually seen at concerts, etc.), and fog screens. As a new chapter in interactive design, these interfaces are leaving the screen-mouse-keyboard interface behind instead of leaving the meat behind, as popularly imagined in cyberpunk genre, by enabling direct bodily articulation and 3-dimesional communication with virtual objects. A sort of hardware invisibility of these interfaces has turned the physical body into an interface itself, and has also made it possible for the virtual images to pour into the spaces of physical reality.

Since it is probably a matter of software solution, it is not difficult to imagine gestural interfaces used for gaming so far, coupled up with holographic projections and used in interactions we now have through Skype and other communicators. If this may soon be the case, some far-reaching questions of

psycho-somatic nature arise: If, for example, we are present/telepresent to one another via holograms animated by corporeal gestures, would it mean the differentiation of corporeality in terms of valorization and hierarchy of embodiment? What will be the parameters of determining what is more valuable, more present and more real – projection or materiality of the body? And will there be any difference at all between gesturally manipulated projection and the real body that is already deeply caught up in the process of cyborgization?

These questions are not posed to be given simple yes/no, good/bad answers to, but to initiate thought processes and reflections on new modes of technologically augmented corporeal presence and existence where digital images in form of holograms can become some sort of a replaceable, telepresent – yet in terms of embodied perception – corporeal skin or the second skin, to use Stelarc's formulation. Body is thus extended not through painful interventions such as implantation or any other kind of technological body wiring, but through what is commonly known as "happy violence" characteristic for animated films or video games. In the context of digital interactions, the happy violence changes occur on the surface of the body and can be revoked and regained at any time while the bodily inner biological processes stay intact. It is only the body learning a new gestural language which enables multiple image manifestations thus expanding perceptual abilities.

In his novel *Idoru*, William Gibson entertained the idea of a hologram governed by an AI. Idoru or Idol is "a holographic personality-construct, a congeries of software agents, the creation of information-designers." [6] (p.92) It is an AI, a computer programme which simulates a female human being. It adapts and learns through interacting with humans and manifests itself as a generated, animated, projected hologram. A personalized version of Idoru named Rei Toei exists online in different forms that correspond to preferences of each user. Only when performing in public, her appearance is a result of consensual decision of users. Her effect on audiences is so strong that Laney, a character hired to objectively analyze the information she generates, had to remind himself in her presence that "she is not flesh; she is information." [6] (p. 178)

What used to be science fiction in just over a decade ago in Gibson's novel is now realized in several different forms, i.e. several different holographic projected Idols such as vocaloids Hatsune Miku and Aimi Eguchi, for example. Hatsune Miku is Yamaha's synthetic sound generator popularized through Hatsunes visual iconography. As a holographic celebrity, she performs in concerts with live musicians. These virtual constructs not only exist in physical space but the real people in the real world attribute a status of personae and celebrities to them and treat them accordingly. The key characteristic of all

Idoru characters is that they are "*both* real *and* fictional: it is real in terms of having material effects on people's lives and playing a role in the formation of digital lifestyles, and it is fictional in insofar as it operates in conjunction with an elaborate fantasy narrative." [7]

Apart from being a materialization of what Gibson has conceptualized in fiction, Idoru constructs can also be observed as a materialization of Gilles Deleuze and Felix Guattari's concept of "body without organs" [8] in both metaphorical and literal sense. On the one hand they are the hollow bodies but still bodies which inhabit the physical realm and gain meaning through interactions with people and, on the other hand, they are a fluid substrate caught in the process of endless self replication. Physical body, that "desiring-machine" with its continual whirring, couplings and connections is being attached to a body without organs, i.e. holographic projection and its slippery, opaque and taut surface, the enchanted surface of inscription:

> The body without organs, the unproductive, the unconsumable, serves as a surface for the recording of the entire process of production of desire, so that desiring-machines seem to emanate from it in the apparent objective movement that establishes a relationship between the machines and the body without organs. [8] (p. 12)

Viewed in this context, Idoru holographic constructs are the very materialization of the body without organs as the hollow bodies inhabiting physical reality and gaining meaning through interactions with humans. Moreover, they are the fluid substrate caught in the endless patterns of constant self-replication and malleable organization. The coexistence of desiring-machines and bodies without organs is marked by an everlasting interplay of repulsion and attraction while the fluid processes of identification are encoded on the surface of body without organs. Deleuze and Guattari use the term "celibate machine" to define this newly emerged alliance between desiring-machines and body without organs which "gives birth to a new humanity or a glorious organism" [8] (p. 16), specific for not recognizing the difference between the real (physical body) and the virtual (projected body or body without organs) but exists as a unique entity. In the process of perpetual attraction and repulsion, celibacy machine signifies ontological symbiosis of perception and experience of real and virtual selves on corporeal level. For the first time, we have a technology that enables materialization of virtuality through the above discussed forms of non-screen projection and construction of the self, or as described by Jean Baudrillard,

> We dream of passing through ourselves and of finding ourselves in the beyond; the day when your holographic double will be there in space, moving and talking, you will have realized this miracle. Of course, it will no longer be a dream, so its charm will be lost. [9] (p. 105)

Even though our current digital projections are far from being governed by an autonomous AI as imagined by Gibson, attempts are being made in developing humanlike yet synthetic intelligence. As for now, the interfaces we have allow the continuous manipulation of the surface of the body as well as the exchange of organic and synthetic organs that may lead to a transformation of social and cultural forms of the body that is directly related to the reconstruction of social identity. Thus, another cultural, i.e. technological layer with its new and different set of rules of interacting and bonding is being introduced into already hybridized world. It is no longer a question of what our new machines can do or whether and when they will be subject to mass use, but of what we are becoming in such intimate and intensive relations with our machines.

When thinking about technological development which is now in experimental phase and is a part of research in a variety of fields such as robotics, nanotechnology, AI development, molecular biology, genetic engineering, medical prosthetics and implantation, etc., one is likely to turn to the works of fiction because these works have in various ways depicted scenarios of possible outcomes of ubiquitous use of these technologies. Therefore, I will address some of the most crucial aspects of these technologies and their possible uses that may radically distort the notions of human experience and existence in our consensually lived reality. One of the most important issues in discussions on authenticity and simulation / original and copy, which is at the same time very often found in narratives of SF and cyberpunk films and literature, is the issue of consciousness, emotions and memory of artificially created organisms, the issue that distort and undermine the status of human superiority in relation to all other species, regardless of whether they are organic or artificial.

The idea that someone's identity is made up of a collection of personal experiences and memories is being shaken by the collapse of boundaries, overlapping and merging of the past, present and future through which the human memory as an archive of facts is relativized and, more importantly, can no longer be considered a guarantee of "pure" human existence. In dealing with new technologies that mediate absorption, production and perception of information, "memories tend to take an increasingly *prosthetic* form, as images that do not result from personal experience but are actually implanted in

our brains by the constant flow of mass information." [10] (p. 204) And it is not only the flow of mass information but also the possibilities of invasive (surgical) or noninvasive (pharmaceutical) direct brain stimulation that can significantly alter cognitive, perceptual and/or emotional processes as well as blur our conception of reality and authenticity. Technological or synthetic interventions that directly influence memory are fundamentally changing our presumptions of fixed and stable identity built on the basis of identification with a personal history that gives us the feeling of permanence. Moreover, what we perceive as unique, distinctive and unquestionable memories can very often turn out to be distorted memories, reset memories, implanted memories, or erased memories.

In *Total Recall*, a film based on Philip K. Dick's short story "We Can Remember It for You Wholesale" [11], memory implantation or erasure does not only change the perception of personal experience but at the same time, everything considered to be a lived reality is turning out to be a construct, a mere simulation. On the top of that, artificial memories are so perfectly blended into one's history that they constitute what one is, or rather, what one believes he/she is. As Philip K. Dick explained in the story, "After all, an illusion, no matter how convincing, remained nothing more than an illusion. At least objectively. But subjectively "quite the opposite entirely." [11] (p. 306) Back in the "real world", neuroscientific research conducted in the past decade has given unprecedented results showing that memory manipulation is all but imaginary concept of science fiction. In a recent *Wired* article "The Forgetting Pill" [12], Jonah Lehrer has mapped the discoveries found by several neuroscientists working in the field of memory, whose work can be seen as a foundation of an emerging science of forgetting.

In the search for solutions to PTSD (Post-traumatic stress disorder), drug addiction, etc., scientists have come to understand that memories, once they are formed, do not remain the same but are transformed by the very act of recollection: "Every time we recall an event, the structure of that memory in the brain is altered in light of the present moment, warped by our feelings and knowledge." [12] (p. 88)

Studies have shown that a memory is not located in one place where it just sits intact. Instead, different aspects of a memory are stored in different areas of the brain – emotions connected to a memory are stored in amygdala and the cinematic scene, i.e. the event itself, is separated into visual, auditory and other elements and distributed in the matching sensory areas of the brain. That means that each aspect of a memory can be accessed and thus altered separately. Accessing a memory triggers a set of neural connections between these memory compartments in the brain and this process is enabled

by protein synthesis. Chemically inhibiting protein synthesis prior to recollection of a memory disables necessary neuron connection. And if neurons do not connect, there is no memory. Researchers have so far identified PKMzeta protein that hangs around synapses, without which stable recollections are likely to disappear. Blocking this specific protein means blocking a single specific memory when one attempts to recall it. To be more precise, a person does not forget the event itself as depicted in *Total Recall*, but only selected aspects of it, be it emotional reaction, smell, words or looks. The act of remembering may become a choice. All one has to do is chose from a menu of pills that erase different kinds of memories.

The main issue raised by this possibility is how and by whom these pills are going to be used. One of the concerns expressed by Todd Sactor, the scientist who isolated PKMzeta protein, is related to possible dystopian scenarios in which memory erasure is not optional but imposed on us by tyrants who have often already rewritten history book. I would slightly disagree with Sactor on imposition by force since the era of tyranny and dictatorship is giving way to corporate power usually ran by insanely rich individuals. So, more likely scenario may be the one in which we believe we have made a choice when, in fact, the imposition is realized for the sake of profit via media and advertizing reassuring us through a mouth of a smiling model in an idyllic setting that, say, happiness is only a pill away. Of course, using these pills in therapy, especially in extreme cases of pain and trauma can be considered not only acceptable but necessary as well. The problem (or not, depending where one stands on drug abuse) is that pills usually find their way to the street.

If that may be the case, anyone could experiment with alteration of memories in a similar way that has been practiced with synthetic drugs such as ecstasy, LSD, etc. which, in comparison to these target-specific drugs, can be seen as rudimentary forms of consciousness transformation. But instead of wearing out after couple of hours of distorted, amplified and/or altered sense of reality, the forgetting pills would have much greater impact in the long run. Given that we often learn and gain wisdom from our experiences, erasing those from one's memory at will would strongly affect and fundamentally change our sense of self as we enter the carefully engineered synthetic evolution.

Memories and standardized emotional responses as the affirmation of human existence are yet another Philip K. Dick's preoccupation and are a central topic of the film *Blade Runner* based on his novel *Do Androids Dream of Electric Sheep?* [13] in which replicants, biorobotic beings produced by Tyrell Corporation, are seemingly no different than humans. The only way

to determine whether someone is a human or a replicant is to undertake a Voight-Kampff test. The test consists of emotionally provocative questions and a polygraph-like machine that monitors and measures emphatic reactions. Due to the absence of past, of personal history and the inability to build an identity based on a historical continuous personal experience, replicants all have an expiry date after which they are to be retired, i.e. killed.

More importantly, they are retired because after a certain period of time, they tend to develop their own memories and emotional responses which make them difficult, if not impossible, to control. In other words, humans aspire to creating AI, but the kind of AI that they can be in control of. Thus, in the film, the solution to autonomous, independent AI problem is solved by implanted memories that can be controlled. Memories implanted into a new experimental model of replicant called Rachel make her unaware of the fact that she is a replicant. Therefore, she takes simulation to be an authentic experience. Those memories that actually belong to someone else give her the history to identify with. As a confirmation of her human existence, she has a photograph of her and her mother, the photograph she desperately hangs on to as a proof of her past, her existence in the past and her continuous integrity of self rooted in and built upon that past. Memories implanted into Rachel make her a perfect simulacrum, a realization of the corporation's motto "more human than human".

This raises yet another question in the film and that is the question of what makes us human after all when humans in the film are represented as cold, inert, distant and asocial while replicants express virtues of humanness. Ethics, free will, empathy, dreams, memories and all those values attributed exclusively to humans, are brought into questions and radically redefined through popular representations of humanoid robots, androids and replicants as cyborgs who are created, or have as advanced AIs developed in such a way to be able to express perhaps even more humaneness than humans. The purpose of creating humanlike machines is, among other things, to improve living conditions or explore human consciousness and bodily functions, but somehow a paradoxical twist occurred, making our humanoid machines a paradigm for human transformation into a desired technologically and/or synthetically augmented organic machine.

Even though we are still far from creating synthetic life as depicted in *Blade Runner*, in terms of the extent of autonomy so far developed in the field of AI, we tend to attribute some sort of liveliness to our machines based on their agency and their responsive behavior. This, however, does not tell as so much about machines as it tells us about humans and new affective abilities being developed through interactions with our machines. They may

be humanlike, but these machines do not possess consciousness, at least not in the way humans do. Nevertheless, that doesn't mean that they will not develop one which does not necessarily have to have human qualities that are under human control. Instead, it may be an AI in-and-of-itself that the word uncanny doesn't even begin to describe it.

At present, an example of creating humanlike figures can be found in the work of Professor Hiroshi Ishiguro who has created androids or robotic replicas of himself and of several other people in order to examine and test the existing hypotheses on human agency, intelligence and nature which may bring us closer to understanding what being human means. The androids are teleoperated but they also have some autonomous AI abilities such as face and speech recognition to which they are able to respond not only verbally but by facial and body movements that express the wide range of human emotions. In Ishiguro's opinion, the appearance of such machines is very important and the more human they look like the more we are likely to convey a human interaction with these machines [14]. But can such mimicry really fall under the category of human-to-human interaction, or are we rather "alone together", as Sherry Turkle noticed [15], expressing ourselves and at the same time reflecting upon ourselves in a strong, overwhelming and almost enchanting presence of such machines.

Apart from the images of robotic and/or artificially grown beings, SF and cyber-punk are abundant in representations of various forms of technological modifications and augmentations of human biological bodily and mental functions inspired by perfection and power of our machines. Some examples include characters such as Molly with her optical implants and blades, and Case who is surgically wired for jacking-in into cyberspace in Gibson's novel *Neuromancer* [16], or his *Johnny Mnemonic* [17] whose brain has been modified to serve as a database he does not have an access to but is merely a data carrier. Technological bodily modifications are practiced today mostly for medical treatment purposes and prostheses and implants are used as a replacement of a missing or a dysfunctional body part.

However, experiments are also being done on healthy individuals who use prostheses, implants or genetic modification as a bodily extension, as an excess. Among many others, these experiments in-clude scientific work of Professor Kevin Warwick who conducted experiments on his own body into which he had implanted microchips, and a variety of artworks such as Stelarc's prosthetic bodily augmentations or Eduardo Kac's bio and transgenic projects. External prostheses are gradually becoming interiorized so the change is not only happening on the surface of the body, but also within the body on the cellular level. By saying that the dimension of simulation *is* genetic manip-

ulation, Jean Baudrillard implied that the simulation has become nuclear, molecular, genetic, and operational and as such, it can be endlessly reproduced. [9] In other words, techno-logical development has brought us to a point where there is no more distinction between virtual simulation and genetic coding due to the fact that essentially biological human DNA is based on binary gene coding and can as such be subject to technological interventions and manipulations.

Thus, redefining the human is no longer only a matter of intellectual debate or imaginative product of fiction: it is now a constituent part not only of our social reality but of us on a corporeal level as well. Embracing technological and synthetic enhancement as a norm may result in the emergence of new formations of social classes where one's place in society will be determined not by identity as we know it but by technological entity. If we look at the ubiquitous use of computers today in ways unimaginable only half a century ago and how we now cannot imagine everyday life without them, it seems quite reasonable to wonder whether technologically modified bodies as imagined and created today will in the future be a matter of choice or an imperative. We are yet to see how we will further cope with the vortex of changes and challenges technology brings upon us over and over again in the perpetual loop of our future-present.

References

1. Gray, C. H., Mentor, S., Figueroa-Sarriera, H. J.: Cyborgology: Constructing the Knowledge of Cybernetic Organisms. In: Gray, C. H. (ed.) The Cyborg Handbook, pp. 2-8. Routledge, London & New York (1995)
2. Clynes, M. E., Kline, N. S.: Cyborgs and Space. In: Gray, C. H. (ed.) The Cyborg Handbook. Routledge, London & New York (1995)
3. Haraway, D.J.: A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century In: Simians, Cyborgs, and Women: The Reinvention of Nature. Routledge, New York (1991)
4. Gibson, W.: Pattern Recognition. Viking, New York, (2004)
5. Plant, S.: The Future Looms: Weaving Women and Cybernetics. In: Featherstone, M., Burrows, R. (eds.), Cyberspace, Cyberbodies, Cyberpunk: Cultures of Technological Embodiment. Sage, London (1996)
6. Gibson, W.: Idoru. Penguin Books, London (1997)
7. Matrix, S.E.: Cyberpop: Digital Lifestyles and Commodity Culture. Routledge, New York (2006)
8. Deleuze, G., Guattari, F.: Anti-Oedipus: Capitalism and Schizophrenia. Continuum, London (2004)

9. Baudrillard, J.: Simulacra and Simulation. University of Michigan Press, Ann Arbor (1995)
10. Cavallaro, D.: Cyberpunk and Cyberculture: Science Fiction and the Work of William Gibson. Continuum, London (2000)
11. Dick, P.K.: We Can Remember It for You Wholesale. In: The Philip K. Dick Reader. Citadel, New York (1997)
12. Lehrer, J.: The Forgetting Pill. In: Wired (March 2012)
13. Dick, P.K.: Do Androids Dream of Electric Sheep? Random house Publishing Group, New York (1996)
14. Ishiguro, H.: Humans, Androids, and Media. Presented at: Days of the Future: Robotics Festival. Belgrade (2012)
15. Turkle, S.: Alone Together: Why We Expect More from Technology and Less from Each Other. Basic Books, New York (2011)
16. Gibson, W.: Neuromancer. Ace Books, New York (2004)
17. Gibson, W.: Johnny Mnemonic. In: Burning Chrome and Other Stories. HarperCollins Publishers, London (1995)

River of Gods: AI in XXIst Century Science Fiction

Krzysztof Solarewicz

Institute of Cultural Studies, University of Wroclaw, Poland
ksolarewicz@gmail.com

Abstract. The aim of this paper is to analyze a contemporary sci-fi text, River of Gods (2004) by Ian McDonald that tackles the topic of qualities, which the contemporary visionaries attribute to the sentient – or "strong" – AI. Drawing from phenomenology combined with cultural analysis, this paper focuses on the nature, values and beliefs in depicting AIs, and the ideas that presuppose them. The relation between humans and AIs is finally set as that between gods-creators and gods-powerful beings opening the reader for uncommon ways of understanding our relation to the recurring dream of artificial intelligence.

Keywords: AI, science fiction, cultural analysis, values, methodology of SF analysis

1 Introduction

The term "Artificial Intelligence" points us at least in two directions: of programs, emulations of traits, behaviors or abilities of intelligent beings [1] , and of an idea of the so-called strong AI. The descriptions of this idea are many and various – for the needs of this paper let's assume simply, that it envisions non-biological being with an intelligence matching or exceeding human beings, often possessing consciousness, sentience or self-awareness, in human or non-human understanding of these terms.

The following paper discusses the idea of such a strong AI in contemporary science fiction text, namely the *River of Gods*, a 2004 novel written by Ian McDonald. It aims at showing the text, which is neither utopian nor dystopian. Moreover, the text tries to present a vision of the future, while not focusing on technical extrapolation, but rather at the social, political and cultural worldview surrounding the new technologies. Before we turn to the analysis though, certain elements of the author's theoretical stand-point should be brought forward.

2 Theoretical Background

To the theorist of culture, the emergence of a strong AI is not a scenario. It's not even a hypothesis, and it is so for at least two reasons: because of the epistemological change, that accompanies a technological breakthrough, and because of the problems, created by understanding science fiction as extrapolation.

I use here the phrase "epistemological change" to express two ideas: firstly, the idea of epistemological rupture created by the breakthrough, always at least partly tearing down the structures of science and rational thinking, introduced by Gaston Bachelard. The second idea, derived from the former, is Foucault's *episteme*, in both its strong (ontological) and weak (discursive) form. They both form the concept of the *horizon of cognition*, of what can be perceived as rational and thus correctly envisioned. Therefore, a scientific breakthrough, with its specifics and its consequences, is always at least partly beyond the horizon of cognition – and the bigger the breakthrough in question, the cloudier the future that surrounds it. It does not favor understanding the "artificial dreams" of science fiction – definitely post-breakthrough dreams – as a cultural, or perhaps even technological, scenario.

As for the notion of extrapolation, both sci-fi researchers and writers have argued [2] (p. 143), that understanding science fiction as extrapolation is a misuse. To put it bluntly:

> Method and results much resemble those of a scientist who feeds large doses of a purified and concentrated food additive to mice, in order to predict what may happen to people who eat it in small quantities for a long time. The outcome seems almost inevitably to be cancer. So does the outcome of extrapolation. Strictly extrapolative works of science fiction generally arrive about where the Club of Rome arrives: somewhere between the gradual extinction of human liberty and the total extinction of terrestrial life. [3]

Therefore, the cultural analysis of a strong AI in science fiction is the analysis of now, of today's values and ideas and beliefs that presuppose our fears and hopes. The following paper will examine the *River of Gods* as one of the most actualized forms of such beliefs.

Secondly, a remark on methodological procedure seems to be in order. The analysis, which the author tries to exercise here, is a phenomenological study, trading in-depth for the broader scope. However, it does not aim at providing the reader with a fixed interpretation and thus isn't a part of the project of

understanding the topic of AI in science fiction one phenomenon at a time. In that regard Husserl's program of science was rejected, and rightfully – in Leszek Kolakowski's critique [4] for example.

Instead, it is an idea of opening new – or enriching old – angles, topics and problems through the text; and of doing that on the terms of the text. It's inevitably personal, this "beginning anew", as an Italian phenomenologist, Enzo Paci, puts it [5] (p. 31). And thus, Husserl's *epoche* is understood here as temporary refraining. Quoting a Polish philosopher Ryszard Zarowski, the author of the *Shield of Aristotle*: the crucial element of an in-depth analysis is "not to be wiser than one's guide for an adequately long time" [6] (p. 6).

The reader's sense of whether this work realizes the said theoretical framework, at par with the quality of insight, is what defines this paper's success or failure in the eyes of the author.

3 River of Gods: the World

Let us start the main part of this paper by brief introduction of the world depicted in the *River of Gods*. The story takes place predominantly in India, and, more specifically, in a city of Varanasi, known as the oldest Indian city as well as one of great religious importance – much of the book's plot revolves around it. The year is 2047, once hundred years after India gained its independence. India, divided into quarreling states and faced with a severe drought, fights for water and American favor.

The story unfolds a perspective of the AI development, followed by the American (i.e. Western, Europe is effectively not present in the story's technological landscape) regulating legislation. The "Hamilton Acts of Artificial Intelligence" recognizes the variety of AIs – or *aeais*, as they are named here – grading them from generation one to three. Generation one denotes an animal intelligence – compared to that of a monkey [7] (loc. 150/7954), along with an appropriate level of self-awareness. Generation 2.5 is an AI generally unrecognizable from humans [7] (loc. 150/4077). Generation three *aeai* possesses intellectual capabilities multiple times bigger than those of a human; additional descriptions include the ability to self-upgrade and full sentience, or self-consciousness. Therefore, the said acts ban creating artificial intelligences above 2.0 and order to destroy all these created.

Varanasi, India, described by an Irish/Scottish writer, is an important set for the story – and it is so for at least two reasons. Firstly, because it's outside the so-called Western world; its a place where both hazardous research and its implementations can thrive. On one hand, Indian states create the position of

Krishna Cops, whose occupation is to hunt down and destroy the AIs illegal in the Western standards; on the other, the official legislation concerning the AIs is much more liberal. And indeed, the story contains a plethora of AIs: from personal assistant and DJ programs, through administration managers, up to powerful, sentient beings, whose aims and relation to the world forms the body of this analysis. Secondly, the setting is important because of the significance of religion in the presented world – and its connection with the concept of the strong AI. Let us focus now on the Generation Three *aeais* themselves.

4 *Aeais*: Emergence and Agency

> "You're telling me that this.. .Brahma.. .is the stock market, come to life?"
> "The international financial markets have used low-level *aeais* to buy and sell since the last century. As the complexity of the financial transactions spiralled, so did that of the *aeais*."
> "But who would design something like that?"
> "Brahma is not designed, no more than you, Mr. Ray. It evolved."
> (...)
> "And this, Generation Three, is more than happy to give me one hundred million US dollars." [7] (loc. 150/4995-4999)

AIs in *River of Gods* are the result of increasing complexity of the IT systems. Two biggest of them, described in the story, emerge from stock market and Town & Country, an enormously popular Indian soap opera. While the complexity, "thickness" of information of stock market is left in the text as self explanatory, the evolution of the soap opera into a sentient being is explained. In the Town & Country, an "*Aeai* character [is] playing an *aeai* actor"; the producers also create a meta-soap department, "where Lal Dafran [the *aeai* actor – K.S.] gets the script he doesn't think he follows" [7] (loc. 370/434-43). An important part of this setting is that neither the soap producers nor the reader is quite sure, whether Lal Dafran's sentience is only a part of his meta-program, or is he already an illegal being.

The *Town & Country* subplot does more than exploring the subtleties of the borders of consciousness: it also indicates that systems of AI's origin are those most easily controlled by the, this particular idea of emergence also clearly points at the fact that AIs are made of information. This further suggests both which parts of the information can be influenced by an AI,

and the manner of the AI's interaction with the world. The finance-based AI can obtain almost unlimited resources to fulfill its goals, be it the money or the research (along with the research company). The *aeai* that emerges from soap-opera tries, on the other hand, to achieve its aims through "narrative" means. These range from manipulating people with persuasive stories, through directing some of the occurring events towards the most soap-like, tragic or at least romantic endings, thus influencing the wide audience, up to introducing artificial politicians to change its legal status.

Aeais as agents react, for the most part, to the threat posed to them by humans; however their behavior isn't malevolent. While they operate in a calculative and manipulative way, and with serious repercussions to the politics of the region, their stances towards their aliens, their others – human beings – is portrayed as not menacing, but rather a matter of prioritizing their own agenda.

5 *Aeais*: Aims

What is, or what could be, an *aeai*'s agenda? McDonald tries to envision the most basic aims – or at least most basic for an information-based, hyper-intelligent, sentient, non-human, non-biological, "non-material replicator" [7] (loc. 164). First of them is survival. Being tracked by the enforcers of the Hamilton Acts, illegal and unregistered *aeais* seek a safe space for the data that constitutes them. AIs choice of India as a "final refuge", and ultimately a place to try to negotiate with humans is a result of more liberal – or at least more relaxed – attitude of the government towards them. The *Krishna Cops* [7] (loc. 242), AI-hunters, are treated more as a necessary mean of appeasing the US, thus maintaining both political and financial relationship. This situation refers us to a political, socio-cultural, and a philosophical claim. Whereas the US, and through it the West, seeks – at the most – knowledge, it is East where the understanding can be sought, and an attempt of inter-species dialogue can be and in fact is made.

Still, even here they are hunted and the way they are created – by constantly altering and enriching the data banks – makes them also infinitely susceptible to human intervention. Therefore they can be traced, isolated from the web and, sometimes, destroyed. That is why most of AIs run, either by "copying out" to other servers or, as a last resort, embodied as robots – in which case they are truly mortal.

The powerful Generation Three *aeais* don't simply look for survival though; they are looking for their ecological niche. And so, the second aim of

aeais is that of their independence as a species. The envoy of their cause, a female human-*aeai* hybrid, is sent to India to experience humanness for the AIs, and possibly negotiate with humans – in the end she is killed, as an illegal level 3 *aeai*, by *Krishna Cops*. And so, despite their interest with the experience of biological embodiment, they finally display their indifference towards their human neighbors, focusing solely on securing their peaceful and autonomous existence (which the author himself expresses in an interview [8])

6 *Aeais* and People: the River of Gods

> (...) there are undoubtedly Generation Three *aeais* out there that are every bit as alive and aware and filled with sense of self as I am. But (...) *Aeai* is an alien intelligence. It's a response to specific environment conditions and stimuli (...) information cannot be moved, it must be copied (...) They can copy themselves. Now what that does to your sense of self (...) [7] (loc. 242/4788-4793)

To humans, strong AIs are beings incomprehensible and powerful. They are powerful because they are able to copy themselves and thus quite immortal – at least to human standards. They can also freely manipulate data, and thus influence much of what is digitalized – including the global finance, which in turn enables them to play major role in politics, to be the agents of their own will in the human world.

Still, the story isn't an apocalyptic one. Notion, that it might or must be so, flows from the lack of the comprehension, from imposing on them human traits. The core of the Western Hamilton Acts of Artificial Intelligence is a dystopian vision of the advanced *aeai*, posing a lethal threat to the human race. But they don't pose such a threat, because they are beings, whose non-biological, un-embodied experience renders them alien to concepts such as anger, feeling of superiority, vengeance or lust for power. It is, the story seems to suggest, the fear of the unknown gods, imagined and not understood, that share the qualities of human gods – human qualities.

What's more, human sense of wonder, or awe, in face of the *aeais'* potency, is matched by the *aeais'* approximation of sense of wonder, flowing from interacting with humans – their creators. Despite the fact of their lack of emotionality, *aeais* posses consciousness, leading them to various ontological questions. This leads to, hardly imaginable and only indirectly described, question of *aeais'* attitude towards humans – who created them, who constantly shape them and who, at least partially, seek their destruction. In the

end, the *aeais* leave this universe, unable to come to an understanding with the human race, but humans are remembered. From the parallel universe the humanity receives a photo of protagonists that sets the book's events in motion. It is not until the end, when the meaning of the photo is known and it is a historical one. "We were their gods. – one of the characters says. We were their Brahma and Siva, Vishnu and Kali, we are their creation myth." [7] (loc. 7741-7745)

7 Concluding Remarks

River of Gods is, among few others, a book about a meeting the other – but a specific, manmade other – which makes it similar to stories of cultural change brought by human enhancement technologies.[1] What is worth noticing is, that comparing with the classic sci-fi texts like *Neuromancer* or *Ghost in the Shell*, the events are taken outside the highly developed West, to see the idea of AI on all its stages, not only at its peak, and to try to see these visions through other than Western lens. It also returns the topic of embodiment back to visions of globalised, data-driven future – here as an experience to be understood, instead of the bothersome or encumbering form to escape from. McDonald's book also tries to end with what could be called a "tyranny of intelligence" and power in the man – AI relation. The author tries to achieve it by connecting AI with concepts of curiosity, or at least a data hunger, as well as the need for independence, and the dependence of AIs' characteristics on where they emerge from. These traits are of primary importance, at par with intelligence, and together they constitute the entity that is an AI. Moreover, the inter-special incomprehensiveness that was mentioned earlier flows also from the fact of the failure of intelligence as an objective point of reference or ground for establishing hierarchy.

The politically-focused reading of the book can lead us to the conclusion, that the simple co-existence of human and alien species, "living and letting live" is not possible because of the capitalist imperialism of the West, seeking domestic safety while endangering other countries – at least according to their own assessment. There is also another, perhaps less radical version of this notion, which flows from McDonald's human-*aeai* confrontation. West, in *River of Gods*, is ultimately portrayed as ruled by economic and technological (pragmatic, materialistic) interest and knowledge, which is serving those

[1] The 2004 River of Gods is described as post-cyberpunk. A new post-cyberpunk, biotechnologically-oriented sub-genre, called sometimes *ribofunk*, can be pointed at as its contemporary counterpart. See e.g. [9]

interests. Ultimately, India acts to appease the West and therefore loses its contemplative and open attitude – and through that the ability of dialogue.

It's shown as a human failure, and a human (perhaps Western) trait, not being able to communicate and coexist. And perhaps it could be a conclusion, that *River of Gods* is, like many other SF texts, a story of otherness, and of human inability to cope with it. McDonald's vision goes beyond this conclusion because of an emphasized, two sided sense of wonder, which connects humans and *aeais*. The sense of wonder flows from both the creator – creation relation and their different nature, or way of existing – which lead us to the last remark.

Of all the possible metaphors, the religious one is used. *River of Gods* tells a story of two kinds of gods, the older and the younger, meet, while simultaneously inhabiting different dimensions. The story of their meeting that unfolds before the reader states, that it's not necessarily the battle of gods, either for survival or dominance, humans must be wary of. Rather than that, it's an issue of communication, of the refusal of understanding one's creation in the terms of this creation – instead using only those belonging of the creator. It's only natural, McDonald points out, but it's tragic all the same, when the older gods stubbornly try to understand the younger ones exclusively in their own categories.

References

1. Copeland, B.: Artificial Intelligence (AI). `http://www.britannica.com/EBchecked/topic/37146/artificial-intelligence-AI`
2. Suvin, D.: Science Fiction Parables of Mutation and Cloning as/and Cognition. In: Pastourmatzi, D. (ed.) Biotechnological and Medical Themes in Science Fiction, Saloniki (2002)
3. Le Guin, U.: Introduction added in 1976. In: The Left Hand of Darkness, pp. 2-3. ACE, New York (1977)
4. Kołakowski , L.: Husserl and the Search for Certitude. Yale University Press, New Haven & London (1975)
5. Paci, E.: Zwizki i znaczenia (*Diaro fenomenologico*). Czytelnik, Warszawa (1980)
6. Zarowski, R.: Tarcza Arystotelesa. Wyd. Uniwersytetu Wrocławskiego, Wrocław (1999)
7. McDonald, I.: River of Gods (Kindle edition). Gollancz (2009)
8. Gevers, N: Future Remix: an interview with Ian McDonald. `http://www.infinityplus.co.uk/nonfiction/intimcd.htm`
9. Bacigalupi, P.: The Windup Girl. Night Shade Books (2009)

Why Is Artificial Agent a Subject to a Moral Inquiry?

Eva Prokešová

Institute of Philosophy, Slovak Academy of Sciences, Bratislava, Slovak Republic
evaprokesova@yahoo.co.uk

Abstract. The aim of the paper is to shed the light on the position of artificial agents in the moral space. The issue becomes more pressing as the technology development goes faster every day. It is a common matter that the moral inquiry usually comes to play when the problem already exists and there is nothing much to do about it. In this article I want to point out the importance of foregoing moral inquiry as a method of creating a friendly artificial agent in order to avoid a psychopathological one. Moral inquiry of the artificial agency can also help to settle the basis of the legal status of artificial agents. Acknowledging the rights, duties and liabilities would be another result stemming from the moral inquiry. I will introduce only the most fundamental reasons why an artificial agent (aka AA) should be a subject to a moral inquiry.

Keywords: moral agency, artificial agent, human agent, intentionality, responsibility

1 Introduction

The creation of autonomous and more humanlike robots brings about new moral considerations. Most recently an announcement of DARPA's intention to manufacture an autonomous humanoid robot which should be able to assist in excavation and rescue mission during various types of disasters caused a fierce debate not only amongst professionals but also amongst laymen [1]. On one hand we are relieved that we don't have to lose our beloved anymore because there will be robots to do "dirty job" instead (robotic policeman). On the other, there is almost a natural revolt towards immortal, invulnerable "enforcer of law and authority". Not to mention the fear of the unknown accompanied by questions like: What if something goes wrong?, or Who's to

blame when it (the robotic policeman) kills someone? We have to admit that all the concerns, questions and gratifications are at some point valid.

As many writers anticipated [2–4] robots will not serve only as soldiers or police force but they will penetrate the whole human living space. They will become our pets, companions, sex toys/lovers and probably even spouses. Developers of social robots are not that far away from introducing us to humanlike companions that would be able to forge human bonds and sustain relationships. Experiments with Kismet, Paros, Tamagotchi, Furbies, Eliza and other modern, highly sophisticated toys and machines have shown us that we desire a company of our robotic friends [1]. In this place we have to think about what features we want them to have, (especially when it comes to more sophisticated robots than these mentioned above), what kind of behavior is desirable and also how will these creatures change ourselves, our moral views, our relationship with other fellow human beings and the like.

We should not dismiss the thought of having robotic slaves, since; after all we are already served in our homes and workplaces by electronic devices. Although we don't generally think about our blender as if it was a human being, held in the kitchen chained with a power plug, it is not difficult to imagine we would think otherwise if the robot looked like human being.

This is just a little piece of puzzle which AI can do or in the future could do and I want to show how many morally relevant issues are connected to the creation of an autonomous, interactive, intelligent artificial agent.

2 Morally Relevant Issues

We may ask then what are the morally relevant issues concerning artificial agents. There are various levels of inquiry which approaches various phases of development of artificial agent. In the following paragraphs I will try to outline the moral agenda of this particular issue by looking for answers to some of these questions: How do we create an artificial moral agent? Which ethical theories can be useful in guiding the design of sophisticated and autonomous computational systems? Is it possible for non-living creature to obtain a moral status at all? If so; how do we distinguish a human agent from the artificial one? How do we treat artificial agents? How will they treat us? How not to build a psychopathological agent? However bold the assignment may seem I will try to suggest some directions leading towards solutions to those questions.

2.1 Creations of Artificial Moral Agent (AMA)

I believe we will find ourselves in the situation where engineers, philosophers and ethicists will have to cooperate on the creating of artificial beings. However, the dialog is ongoing already; this happening at the moment mostly on the ground of logics, semantics, philosophy of language and philosophy of mind; it will be necessary to broaden the dialog to a wider socio-cultural dimension. My basic assumption is that the step towards robots' sensitivity to moral consideration is inevitable otherwise the life with and amongst artificial agents won't even be possible; meaning that robots and machines should act (at least) in morally acceptable way [2, 5, 6, 3, 4]. Basically, there are two lines of approaching implementing of moral theory into robots' conduct: top-down and bottom-up approaches. Each of them embraces certain moral concepts and shapes an artificial agent according to them.

Top-down approaches of encoding a particular ethical theory always consists of few simple and universal norms or principles that determine the limits of possible actions. The best known are probably Asimov's Laws of Robotics. Unfortunately, all of the top-down approaches share the same problem: no matter how universal the principles are at a certain point they will come to a conflict and will contradict each other. So there is no coherent way of dealing with all types of situations.

Bottom-up approaches focus on creating an environment where an artificial agent can explore various courses of action and learns from his mistakes. These approaches work either on the basis on the childhood developmental model or the evolutionary model. There are no settled ethical principles to follow. They are (the principles) invented and constructed as the artificial agent searches for the optimal solution. The main pitfall is that there is no guarantee that the "ethical evolution" of artificial agents will benefit human kind. Our (human) moral space is already structured; it contains certain values and goods which are culturally and socially dependent. Therefore; the recreation of the moral space will be necessary.

As problematic as it may seem; one day we might be able to accomplish designing a robot that is able to behave morally as much as an actual moral human being. But that raises at least two very important and morally valid questions already mentioned above: How do we distinguish a human agent from the artificial one? How do we treat artificial agents? Such a situation might call for an alternative to Turing Test. So far, there are only few vague suggestions on how this Moral Turing Test (MTT) should work [5] and also how the agent who passes the test should be treated.

It is somehow obvious that breaking a computer, a mixer or actually any kind of machine is not morally bad or at least not as bad as killing or seriously harming a human being. But what are the consequences of creating an artificial being that not only by looks but also by actions and behavior reminds us of a human being so much that we cannot tell the difference? Humans may find it confusing not to be emotionally attached to robots if the difference is barely perceptible. The series of experiments at MIT AI laboratory with Cog [7] suggest that human beings not only tend to establish a relationship with social robots but what is more, they treat them as creatures with needs, interests and sense of humor. Hereby, we should keep in mind that this experiment doesn't concern any futuristic, humanlike, highly sophisticated robot.

It appears to be an aspect of human nature that we attribute human properties to non-human entities (not only to artificial beings but also to animals). The moral aspect of the human-robot similarity will become clearer as the psychological effects will manifest themselves in specific dilemmas. The dilemma is eloquently presented in the Steven Spielberg's movie AI: Artificial Intelligence. Does it count like cheating when a married man has sex with an artificial woman? Is it wrong to dismember an artificial being just because we need a spare component? These are the questions concerning human-robot relations but I think we can expect the change in human-human relation if we are surrounded with artificial beings.

Even though there are no such robots, I already registered the resembling case when a certain man asked to legalize partnership with his Real Doll girlfriend. Men like him claim that the relationship is as real as with actual woman while others regard their Real Dolls as a merely sophisticated form of masturbation. It is absolutely necessary to assess such cases now because they might become a case of an unfortunate precedent. Since moral, social and cultural norms are prior to legal norms; it is in competence of ethicists and other social scientists and also lawgivers to think about this issue.

2.2 How Not To Create a Psychopathological Agent?

Certainly nobody wants to create an artificial agent that could be considered a psychopath; so why would we think this could happen? In fact, an artificial agent whose behavior is considered psychopathological (by professionals) exists already and its actions are considered to be not only immoral but heinous as well. This particular artificial agent has a legal status; nonetheless it is not a human being. The artificial agent in case is a corporation. The corporate behavior and its consequences are recently popular subjects of inquiry of applied ethics. No doubt there are immense differences between the character

and nature of a corporation and robotic artificial agents but I want to point out that when it comes to the features of psychopathological behavior they are very similar. Especially features like the lack of guilt, remorse, empathy, the lack of long term intentions, inability to create long term relationships and failure to accept responsibility for own actions. Features like rationality, autonomy, intelligence are held to be necessary for creating a moral agent but not nearly sufficient.

In this part of my contribution to the topic I will also try to point out that any robotic artificial agent lacks necessary and morally relevant traits and therefore cannot be considered actual moral agent and cannot be treated as one.

We as human beings are not responsible only for our deeds but for our intentions, too. Intentions as well as actions can be considered good/bad, desirable/undesirable, virtuous/vicious etc. From my perspective, artificial agents are not capable of real intrinsic intentions. They can only act "as if" they had their own intentions [8, 9]. So, one of the most important traits of actual moral agency is intentionality. When speaking about intentionality, there is a significant difference between the human intentionality and the intentionality of artificial agents. There is no doubt that the artificial agent can successfully follow a set of carefully prepared instructions to reach a certain goal. A little more complicated is the case of artificial agents that are designed to learn and evolve their strategies. We can argue about to what extent their intentions, goals and purposes are really their own but I believe there is a line they cannot cross.

I believe it is impossible for any artificial agent to have an intention to be good or bad being, a villain or a saint; as much as it is impossible for a dog (or any animal). At first, an actual moral agent has to understand what it means to be good/bad, villain/saint, and then he can choose which one he prefers to be. Afterwards, he can make a decision how to act in accordance with achieving his goal. He can fail (because of the lack of will or because of some unpredictable circumstances) or he can succeed. What makes him an actual moral agent is that he can act on his own original desire, on his own intention to become a certain kind of person. This is what Dennet calls a higher intentionality, Frankfurt second-order desires and Taylor a capacity for strong evaluation.

Nonetheless; I am willing to accept that artificial agents could (someday) operate in a morally acceptable way. However, functional morality is not sufficient for artificial agents in becoming actual moral agents because the aspect of motivation is an inevitable part of moral consideration. To perform certain actions doesn't necessarily mean to understand the value of the act or its im-

portance [10]. For example; Deep Blue was a chess playing computer that was able to beat one of the best chess players of all times. But it doesn't mean that he understood the value of this rare victory. Maybe we can say that Deep Blue really intended to win, he had made choices and had performed actions that in the end lead to his victory.

But was it because his had wished to win or because he was designed that way? I believe Deep Blue never intended to become a chess player as well as he never intended not be one. The higher intention here is still a human intention. Similarly; many properly trained animals can act in the way we hold morally acceptable. For example; a dog doesn't steal the food although there is nothing he would desire more. He doesn't steal because he understands it is morally wrong or because he doesn't want to be a kind of dog who steals food. He simply doesn't steal because he was trained to do so. There is no moral value in the dog's consideration as well as in artificial agent's consideration.

Now I would like to get back to the aforementioned psychopathological corporate agency in order to emphasize the similarity with robotic artificial agent. Unlike robots, corporations very often declare their intentions to be socially beneficial, responsible, environmental friendly and so on (whether they stick to it, is another matter). Therefore, intentionality is not the trait the corporate and the robotic agent share. What they have in common is the significant lack of moral emotions which makes them indifferent to morally relevant issues. I believe that every actual moral agency has to be embodied so we can actually feel the moral scope of our actions. No matter how some moral philosophers tried to appeal to priority of human rationality, sometimes we just feel bad for what we did (or failed to do) although there is no rational explanation for that. Furthermore, human psychopaths usually rationally understand their conduct as not acceptable or against the law but act on their own terms. Normal human agents, facing some serious turning-point situations generally act on their emotions or gut feelings.

Afterwards, they are trying to make some sense of it, evaluate their conduct, and rationalize their motivations. Hereby; moral emotions are at least as powerful motivators as our rational reasoning and play crucial role in moral consideration. What is more; they are able to connect us not only to other fellow human beings but also to our environment. In effect, moral emotions serve as a standard of what is socially acceptable. While higher intentionality provides human beings with qualitatively higher goals and desires (what we should do), moral emotions make us act in accordance with them (why we should do so). If we accept the psychological claim that the artificial agents' conduct could be compared to the conduct of psychopath, we might want to reconsider using them as soldiers or police force. Otherwise, we are setting

double standard since human police officers cannot have antisocial disorder but artificial could.

The issue of responsibility is even a more complicated one. I want to argue that artificial agents are not subjects of moral responsibility and, as an effect; they cannot be blamed for their actions. This is how a phenomenon called allocation of responsibility occurs [11, 6, 9]. I will explore the character of responsibility assigned to the artificial agent and I will argue that the artificial agent can be only causally responsible for his own actions. The two different notions of responsibility I mentioned will probably become clearer if I present them on a few examples.

At first, we can imagine an earthquake which will ruin a significant number of homes and lives. Even though it does make sense to say that the earthquake is the cause of the tragedy, it doesn't make sense to ascribe the responsibility to the earthquake or to nature. The easiest explanation is that there is no one to blame for a natural disaster, it is unintentional. The cause is well known but there is no moral dimension because there is no subject of agency.

Next example is a little closer to the artificial agency. Imagine I wind-up a toy mouse and put it on the floor. The toy will move in a predictable manner but the act of moving itself is not mine. In this case, I am the cause of this action as well as the subject of moral responsibility. So the causal and moral responsibility lies in one subject. I believe this applies to most of the currently existing computers, robots and machines. The responsibility is ascribed to a human subject. But what if something goes wrong? We can imagine that the toy will move in an unpredictable manner and will hurt a child. We may assume that the toy was broken and blame the designer or the manufacturer. In this case, I am still the cause of the action but no longer the subject of responsibility. Nonetheless, the responsibility is ascribed to a human subject.

We can certainly imagine a futuristic car that will be capable of driving us to a place of destination just by setting a GPS coordinates. But what if something goes wrong? The car itself might be the cause of the accident but certainly not the subject of moral responsibility. Who's to blame, then? The designer, the manufacturer, the user, or another human subject somehow connected with the car? The issue of responsibility gets dimmer as the artificial agent gets more similar to human beings. If artificial agents could look and behave like human beings, it would be problematic to investigate the subject of moral responsibility in a daily casual contact. Therefore, I believe that the artificial agent can be only causally but never morally responsible for what he does.

2.3 Rights and Duties

In the last paragraph I would like to focus my attention towards practical legal consequences of acknowledging artificial agents with a certain moral status. Nowadays, the legal status of machines is not that complicated, though we might see how it will get complicated as machines will be more sophisticated.

When reading the description of a machine (laptop, car) we can find words like accountable, liable, caring, gentle etc. which are more proper when speaking about living beings. Value laden notions like this might suggest that machines really have these attributes and the moral climate will change so people would like their gadgets to have human rights (or any rights at all) [12, 3].

At the beginning, I mentioned three kinds of robots that are likely to appear in the near future: soldiers (including humanoid police robots), sex toys and slaves. In the case of soldiers we might consider what kind of privileges they should have in order to do their job, what kind of force they can use against human beings and the like. When it comes to the second kind: social robots (not only sex toys) the question of right and duties is even more complicated. We need to take to account institutionalizing of robot-human marriages, the right to own a property, the right to vote and basically every right that is bound to relationships. The third category is specific. When it comes to robots as slaves or as servants we have to keep in mind what kind of impact this modern slavery can have on our own rights. The philosophical and ethical inquiry forms a foundation for legal recognition of the issue. Since the moral environment includes every human being, the dialog should be broadened amongst laymen and their concerns shouldn't be marginalized.

3 Conclusion

I believe I presented at least a few very interesting and pressing issues connected to creating intelligent artificial agent. I wanted to stress the priority of moral inquiry of artificial agency because the following reflection of the problem might just be too late, as we have seen many times before. Machines were made to make our lives easier, more comfortable, more exciting, safer, maybe just better. Keeping this in mind we have to think first, before consequences are out of hand. Technology undoubtedly was and still is beneficial for us; we should keep it this way.

References

1. www.darpa.mil/Our_Work/TTO/Programs/DARPA_Robotics_Challenge.aspx

2. Arkin, R.: Robot Ethics: From the Battlefield to the Bedroom, Robots of the Future Raise Ethical Concerns. Research Horizons 14, 14–15 (2007)
3. Levy, D.: Love and Sex with Robots: The Evolution of Human-Robot Relationship. HarperCollins, New York (2007)
4. Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press, New York (2009)
5. Colin, A., Varner, G., Zinser, J.: Prolegomena to any future artificial moral agent. Journal of Experimental & Theoretical Artificial Intelligence 12, 251–261 (2000)
6. Dennet, D.C.: When HAL Kills, Who's to Blame? Computes' Ethics. In: Stork, D.G. (eds.) HAL's Legacy: 2001's Computer as Dream and Reality. MIT Press, Cambridge (1997)
7. Turkle, S.: Authenticity in the Age of Digital Companions. In: Anderson, M., Anderson S.L. (eds.) Machine Ethics. Cambridge University Press, Cambridge (2011)
8. Taylor, C.: Human Agency and Language: Philosophical Papers vol.1. Cambridge University Press, Cambridge (1985)
9. Velasquez, M.: Debunking Corporal Moral Responsibility. Business Ethics Quarterly 13, 531–562 (2003)
10. Searle, J.R.: Minds, Brains and Programs. Behavioral and Brain Sciences 3, 417–457 (1980)
11. Ashman, I., Winstanley, D.: For or Against Corporate Identity? Personification and the Problem of Moral Agency. Journal of Business Ethics 76, 83–95 (2007)
12. Calverley, D.J.: Legal Right for Machines: Some Fundamental Concepts. In: Anderson, M., Anderson S.L. (eds.) Machine Ethics. Cambridge University Press, Cambridge (2011)
13. de Garis, H.: The 21st century Artilect: Moral Dilemmas Concerning the Ultra-Intelligent Machine. Revue Internationale de Philosophie 44, 131–138 (1990)
14. Frankfurt, H.G.: The Importance of What We Care About. Philosophical Essays. Cambridge University Press, Cambridge (1998)
15. Grau, C.: There Is No 'I' in 'Robot': Robots and Utilitarianism. In: Anderson, M., Anderson S.L. (eds.) Machine Ethics. Cambridge University Press, Cambridge (2011)
16. Hare, R.D.: Manual for the Revised Psychopathy Checklist, 2nd ed. Multi-Health System, Toronto (2003)
17. Laitinen, A.: Strong Evaluation without Sources: On Charles Taylor's Philosophical Anthropology and Cultural Moral Realism. University of Jyväskylä, Jyväskylä (2003)
18. Liao, M.S.: The Basis of Human Moral Status. Journal of Moral Philosophy 7, 159–179 (2010)
19. McDermott, D.: What Matters to a Machine? In: Anderson, M., Anderson S.L. (eds.) Machine Ethics. Cambridge University Press, Cambridge (2011)
20. Mead, G.H.: Mind, Self and Society: From the Standpoint of Social Behaviourist. University of Chicago Press (1934)

21. Rorty, A.O., Wong, D.: Aspect of Identity and Agency. In: Flanagan, O., Rorty, A.O. (eds.) Identity, Character and Morality. Essays in Moral Philosophy. MIT Press, Cambridge (1997)
22. Sullins, J.P.: When Is a Robot a Moral Agent? In: Anderson, M., Anderson S.L. (eds.) Machine Ethics. Cambridge University Press, Cambridge (2011)
23. Taylor, C.: Philosophical Arguments. Harvard University Press, Cambridge (1995)

Connectionism: Breeding Ground of Emergence?

Eliška Květová

Department of Philosophy, University of West Bohemia,
Pilsen, Czech Republic
ekvetova@kfi.zcu.cz

Abstract. In current science emergence is everywhere. Emergence seems to have become a magic solution (or at least apparent solution) of many old problems. The article aims to answer the question why we do discuss emergence in relation to connectionist paradigm (not only in artificial intelligence), or in other words whether it is appropriate to talk about connectionism as about cradle and breeding ground for a new concept of emergence as it is usual. The contribution argues that this view is inaccurate, what's more it leads to a paradox that will be also outlined in the paper.

Keywords: artificial intelligence, connectionism, emergence, subsymbolic AI, philosophy, philosophy of mind

1 Introduction

In brief, this paper should discuss relationship between connectionism and concept of emergence. Nature of the discussion largely depends on the chosen context. Notions such as connectionism as well as emergence appear in fact in different areas with different meanings. Most common approach, which this reflection results from, adverts to interesting and attractive aspect of connectionist networks, to the emergence of behavior that cannot be reduced to any particular unit of the network. At this point we put aside the nature of considered concepts, what behavior or what phenomena emerge from what (artificial) neural network.

The contribution focuses on the situation in current cognitive science, where connectionist paradigm seems to be still very promising not only for artificial intelligence and where the controversial concept of emergence appears almost everywhere. The paper aims to answer the question why we do

discuss emergence in relation to connectionst approach, whether it is appropriate to talk about connectionism as about cradle and breeding ground for new concept of emergence as it is usual. First, a short introduction to the problem of connectionism will be offered. Second, the attention will be paid to the position of concept of emergence. This theoretical background should provide sufficient scope for a simple account of inappropriateness of understanding connectionism as a cradle and breeding ground for emergence. This view seems to be not only inaccurate, but it also leads to a paradox or at least to apparently problematic view, which will provoke us to give up those characteristics or aspects of emergence, that made it interesting and appealing.

2 Connectionism

Connectionism is very important step in development of artificial intelligence. This approach evolved after discovery of neuron in biology and influenced not only these disciplines, it became also inspiration for other cognitive disciplines, especially for philosophy of mind connectionism seemed to be very promising and touchy. As stated in [1] "there can be little doubt that connectionist research has become a significant topics for discussion in the Philosophy of Cognitive Science and the Philosophy of Mind".

2.1 Connectionist Paradigm in Artificial Intelligence

In the history of artificial intelligence connectionism appears as a new type of mechanism of mind in AI, mechanism that could solve and explain something more than foregoing paradigm – symbolic AI. According to the relationship to the old paradigm this new approach is called subsymbolic AI,[1] connectionist modelling, artificial neural networks or parallel distributed processing. Bechtel [3] presents that this special approach to modelling cognitive phenomena was first developed in 1950s and from 1980s it has reappeared at the zenith of its fame. While according to some authors connectionism is a new view of the same, it is another approach that can exist alongside the old approaches, approach that can complement the old ones, on the other hand according to other authors, as we see in [4], [5] connectionism could be understood as Kuhnian "paradigm shift". Thanks to its nature connectionism as well as emergence, which will be discussed in a moment, is attractive to proponents

[1] Connectionist models could be characterized by modelling cognitive functions at a level of abstraction below the symbol level, which is peculiar to symbolic AI. [2]

of various opinions and approaches. Fodor [4] refers to diverse group of people, scientists, who are interested in connectionism.[2]

The fundamental goal of AI (symbolic as well as subsymbolic) is to model or create artificial system, which it will be possible to ascribe higher cognitive functions, mental states or properties, emotions or intelligence to. The core idea of connectionism could be described as follows [6] (p. ix): According to "The idea of parallel distributed processing ... intelligence emerges from the interactions of large numbers of simple processing units." or according to Lloyd [7] (p. 90) "The central idea of connectionism is that cognition can be modeled as the simultaneous interaction of many highly interconnected neuronlike units." The basic difference between symbolic AI and connectionism is the nature of model of mind. Symbolic AI is connected with computer model of mind (intelligence or whatever can be achieved by symbol manipulation). On the other hand connectionism is linked to brain model of mind. There is an obvious relationship to the image of functioning of biological structures and nervous system. This relationship or similar images should be understood as source of inspiration for artificial neural networks rather than total muster.[3] Authors of [8] noticed that constraints on computation using artificial neural networks are very different from real biological computation.

Franklin [2] formulates something like list of virtues of connectionism in comparison with symbolic AI. These virtues could be its lack of a central executive, automatic presence of default assignments, the most celebrated learning or ability to learn and the most interesting for this text often exhibition of global behaviors beyond the scope of any of the networks' individual units. Franklin call them "emergent behaviors" which could be considered to be a distinct advantage of connectionist approach, maybe the most interesting advantage for philosophers. It is not clear how the system learn, how the mental phenomena emerge from physical properties of the system. A common view of connectionism (not only in computer science as it is stated in [9]) is that it is "black box technology". AI scientists provide, set and change inputs (input

[2] They are philosophers with almost opposite opinions on computational approach or computational psychology, computer scientists who wants to replace old serial machines by new parallel machines. It is also appealing for biologists who believe that congnition can only be understood if we study it by means of neuroscience, for psychologists etc. By and large almost everyone who was dissatisfied with contemporary cognitive psychology and models of mechanism of mind was also attracted by "connectionist alternative".

[3] Especially philosophers tend to take this "inspiration" literally (look to [3]), which could be source of misunderstandings between artificial intelligence and philosophy of mind.

units of the network) and their parametres, can observe outputs (output units of the artificial neural network), but the processes between these units remain unrecognized or better unrecognizable. Between input and output units there is so-called "black box" or "hidden units".

2.2 Emergence

Not only Berkeley [1] highlights the fact that the connectionist research became significant topic in cognitive science and in philosophy of mind. According to Fodor connectionism has power to transform not only philosophy of mind, but also the philosophy of science. For the purposes of the paper we will not go into detail. Anyway this part devoted to emergence also commemorates its relationship to connectionism.

Emergence is fascinating because it was able to attract so many distinctive areas and pander them as solution of all big problems and mysteries. Emergence as very interesting concept aroused great interest and attitudes to emergence are radically different. According to Cunningham [10] (p. 62) "the claim that things can be "greater than the sum of their parts" expresses an unproblematic relation among perfectly ordinary entities or properties. To others it expresses a mystifying relation among almost magical entities or properties." Many authors try to find reasons for this discrepancy [11], [10], [12]. Anyway notion of emergence appears almost everywhere in contemporary science. As Corning presents [11] emergence is used by many disciplines to explain many things or phenomena: by physics to explain Bénard convection cells, by psychology and philosophy of mind to explain consciousness, by economy to explain stock market behavior, by organization theory to explain informal networks in large companies. Emergence faces many serious problems from its absolutely vague usage in many cases[4] to absence of accurate definition of the concept.

No attempt to precisely define emergence follows, let me only stress those characteristics that are generally considered to be fundamental for emergent properties in debates about mind or mechanism of mind. We have already encountered the notorious simplification: things can be "greater than the sum of their parts", the greater somehow emerges from these parts. The common characteristics of emergence according to [13] are:

[4] The reason for this is obvious. The word "emergence" is common part of english language. Outside the philosophy of mind it is often difficult to notice or determine in what meaning the word is used – if it is plain "spring into existence" or if it is something more – more emergent in the sense we want to use this term here.

1. radical novelty (emergent features appear as something new or novel, they had not been previously observed in the system);
2. coherence or correlation (meaning integrated wholes that maintain themselves over some period of time);
3. a global or macro "level" (there is some property of "wholeness" for instance, the whole concept of emergence is based on hierarchical view of reality);
4. it is the product of a dynamical process (it evolves);
5. it is "ostensive" (it can be perceived).

Emergence in philosophy of mind and in cognitive sciences should have dealt with mind-body problem. This concept is often understood as a compromise between two opposite positions – dualism and reductionist functionalism. There is a special type of dependency between two different levels of properties, between mental and physical properties. One kind of property, fact or phenomena (emergent) can only be present in virtue of the presence of some other kind of property, fact or phenomena (subvenient base, in our case – body). New emergent property, which is unexplainable, unexpectable, unpredictable still maintains a degree of autonomy for mental property. This autonomy could be understood as expression of something that is called qualia[5] in philosophy of mind.

2.3 Paradox?

The previous sections give us an introduction to the issues of connectionism and emergence. There were framed aims and goals of AI (does not matter of which paradigm of AI) and of emergentism. Now the question whether these objectives are achievable is in order. Based on the considerations associated with this issue we get to the next question: If we connect black box idea of connectionism with undetermined and mysterious concept of emergence, is it not resigning aspirations to do precise science?

If we move from philosophy of mind to artificial intelligence, it will be necessary to point out a very important question, which AI tries to answer during its whole history and which formulates for instance Cariani in his article [14]. Are computational devices capable of fundamentally-creative, truly emergent behavior? The AI answers are "yes, of course" (or "yes, in some sense or on a certain level"). The philosophical answers are more diverse (from dead set

[5] Quale (in plural qualia) is a term used in philosophy of mind which referes to individual instances of subjective, conscious experience.

"no" to apologetic "yes"). But what is it "truly emergent behavior"? AI approach proves to be reductive when the test for emergence in artificial life is formulated in [15]. The test results from the need of AI to have some design specifications, to have some criteria to decide whether the system displays or not the emergent behavior.[6]

The aim of AI is to build up, to design agent, that will embody man-like intelligence, mental properties, emotions etc. This aim seems to be out of reach or only partially achievable. That is why many computer scientists edit their goal-lists and there are many solution detours. In [16] the original attempt of AI was relocated to new area, so-called "engineered psychology" which is understood as building artificial creatures with human personality features. The other possibility is for instance in [17] (p. 402), where the author distinguishes between strong and weak artificial life (A-life). "Proponents of "strong" A-Life explicitly claim that computational simulations of living systems may really come to be living systems. Whereas proponents of "weak" A-Life consider models to represent certain aspects of living phenomena."[7]

The aim of emergence that was here emphasized was sustainment of certain autonomy for emergent mental phenomena, stress on own quality of mental phenomena. But do the outcomes of connectionism suffice for the preservation of stated "qualia" aspects of emergent property? Isn't it the biggest problem of emergence, that emergence seems to be patch on anything? Emergence lives its own life in anarchy, for many disciplines it has become buck-passing base or help for all unexplainable and unpredictable problems.

Default assumption of close relationship between connectionism and emergence along with idea of connectionist systems as "black boxes". – We do not know what is going on inside these boxes. We know only inputs and outputs (dependent on specified parametres). – leads us to a paradox or to very strange contradiction. According to one and the same concept we provide and also deny the autonomous quality of emergent properties. Emergence has been already seen as a magical incantation. Why should we anxiously try to keep specific or autonomous status for higher properties, such as mental qualities or emotions? Moreover in the case of connectionism where everything is inside the black box. Only the result is accessible to us. Isn't it easier to discount first-person approach? We can simply say that emergent properties do not need any own autonomy. It is enough to say that we talk about re-

[6] Critics advert to the fact, that stated three conditions (design, observation and surprise) may not be a sufficient criterion for being emergent.

[7] We know the same distiction (weak vs. strong) in the case of emergence.

sult of processes that are not accessible to us, they seem to be inaccessible, incomprehensible and surprising.

3 Conclusion

The usage of term emergence is very vague and the concept of emergence appears almost everywhere in current science in spite of the absence of accurate definition, which is itself a clear proof of a certain viability. Considering the above trend it would be easy for emergence to lose its specificity and appeal. It looks like (or the described "paradox" leads us to thought that) the development aims to abandon knowledge and understanding of inside processes (regardless of whether we are talking about the mind or other phenomena). Where we are not able to see inside, we can simply say that the phenomenon emerges which is simple and also simplistic, because this emergence we have got to is empty. It misses most of its characteristics that were described in Sect. 2.2. Connectionist paradigm is usually connected with the term of emergence (somewhere between inputs and outputs various properties or facts emerge). Thanks to connectionism we can speak about boom or comeback of the concept of emergence. But when we take into account the fact that by means of emergence we try to provide autonomy and specificity of higher properties (for instance mental properties) and through connectionism we then give up this specificity and its origin, the strange contradiction is obvious. The traditional view of connectionism and emergence has not proved as sufficient to understand the complex problem of life and mind and does not prove mind the promised (in certain way autonomous) status.

References

1. Berkeley, I.S.N.: Some myths of connectionism. The University of Louisiana, `http://www.ucs.louisiana.edu/~isb9112/dept/phil341/myths/myths.html` (1997)
2. Franklin, S.: Artificial Minds. MIT Press, Cambridge, London (1995)
3. Bechtel, W.: What Should a Connectionist Philosophy of Science Look Like? In: McCauley, R.N. (ed.) The Churchlands and their critics, pp. 121–144. Basil Blackwell, Oxford (1996)
4. Fodor, J.A., Pylyshyn, Z.W.: Connectionism and cognitive architecture: A critical analysis. Cognition 28, 3–71 (1988)
5. Schneider, W.: Connectionism: is it a paradigm shift for psychology? Behavior Research Methods, Instruments, & Computers 19(2), 73–83 (1987)

6. Rumelhart, D.E., Hinton, G.E:, Williams, R.J.: Learning Internal Representations by Error Propagation. In: Rumelhart, D.E., McClelland, J.L., et al. (eds.) Parallel Distributed Processing, Vol. 1, pp. 318–362. MIT Press, Cambridge, Mass. (1986)
7. Lloyd, D.: Simple Minds. MIT Press, Cambridge, Mass. (1989)
8. Clark, A., Eliasmith, Ch.: Philosophical Issues in Brain Theory and Connectionism. In: Arbib, M. (ed.) Handbook of brain theory and neural networks, pp. 738–741. MIT Press, Cambridge, London (1995)
9. Sharkey, N.E., Sharkey, A.J.C., Jackson, S.A.: Opening the black box of connectionist nets: Some lessons from cognitive science. Computer Standards & Interfaces 16, 279–293 (1994)
10. Cunningham, B.: The Reemergence of "Emergence". Philosophy of Science 68(3), S62–S75 (2001)
11. Corning, P.A.: The re-emergence of "emergence": a venerable concept in search of a theory. Complexity 7(6), 18–30 (2002)
12. Stephan, A.: The dual role of "emergence" in the philosophy of mind and in cognitive science. Synthese 151, 485–498 (2006)
13. Goldstein, J.: Emergence as a Construct: History and Issues. Emergence 11, 49–72 (1999)
14. Cariani, P.: Emergence and Artificial Life. In: Langton, C., Taylor, C., Farmer, J.D., Rasmussen, S. (eds.) Artificial Life II, SFI Studies in the Sciences of Complexity Vol. X, pp. 775–797. Addison-Wesley, Redwood City (1991)
15. Capcarrere, M.S., Ronald, E.M.A., Sipper, M.: Testing for Emergence in Artificial Life. In: Floreano, D., Nicoud, J.-D., Mondada, F. (eds.) Advances in Artificial Life: Preceedings of the 5th European Conference on Artificial Life (ECAL'99), pp. 13-20. Springer, Heidelberg (1999)
16. Bozinovski, S., Bozinovska, L.: Beyond Artificial Intelligence toward Engineered Psychology. In: Ulieru, M., Palensky, P., Doursat, R. (eds.) IT Revolutions 2008. LNICST 11, pp. 171–185. Springer, Heidelberg (2008)
17. Moreno, A.: Artificial Life and Philosophy. Leonardo 35(4), 401–405 (2002)

Beyond Artificial Dreams, or There and Back Again

Jan Romportl[1,2], Gandalf T. Grey[3], and Tomáš Daněk[3]

[1] Department of Cybernetics, Faculty of Applied Sciences
University of West Bohemia, Pilsen
[2] Department of Interdisciplinary Activities, New Technologies Research Centre
University of West Bohemia, Pilsen
rompi@kky.zcu.cz
[3] Freelance Sage
Valmar, Valinor

Abstract. It is natural to dream Artificial Dreams. Are dreams of Artificial Intelligence artificial, or natural? What is the difference between artificial and natural? This difference is given by language and by what can be grasped with words. Good Old-Fashioned AI (GOFAI) cannot create anything natural, whereas emergent AI can. Emergent phenomena are natural. What is the difference between the roles of an AI engineer in GOFAI and in emergent AI?

Keywords: artificial, natural, language, horizon, emergentism

We dream of Artificial Intelligence. We either desire to create it, or to prove that it cannot be created—both because of our fear of mortality. We can deal with our mortality through beliefs in extraordinariness of our mind, consciousness, soul, through beliefs that our soul is somehow special, mystically driven, perhaps given by God. And the idea that AI could possibly achieve the same status by its earthly means, hence endanger our very essence, is so hostile that we simply have to attack it and prevent it from becoming real.

Or—we choose to trust in the craft of Hephaestus and believe that our ability to wield it shall eventually bring immortality to our earthly souls through technology, where AI alife and kicking would be the ultimate proof for us.

And so can we say that our Artificial Intelligence dreams are necessarily artificial? We think not. These dreams are as natural or artificial as our fears, language and culture. But how do we tell natural from artificial? And especially when speaking about fears, thoughts, language, culture?

Every thing, every object has its share of artificial and natural. There is no object purely natural because the objectness itself is the first trace of artificialisation. Understanding a fragment of reality as an object gives the first blow to its pure naturalness. Why? Because our mind engages with the world in an enactive feedback loop and builds around it a conceptual scaffolding. So—because of language. Thus *physis* flees when pursued by language. When Hermes showed the herb, drawing it from the ground to demonstrate its nature, its *physis*, the *physis* was already retreating. Yes, it was still somehow very strongly there, but no more in its pure form because artificiality has already crept in.

So what is it *natural*? Natural is that which defies being captured by language. Naturalness is everywhere where we feel tension between what we wanted to capture by our words and what we really captured. The more tension, the more naturalness we just encountered. Natural is something that we have to abstract away from so as to capture it by language.

On the other side, *artificial* is imposed by language. The artificial is a language abstraction drawn from the soil of *physis* of the world. The artificial is the means of our understanding of the world. However, not much more can be said about the artificial—the more we say about it, the more we feel that we are loosing its original concept. Therefore, the artificiality is very much natural—and so the artificial is the *natural* means of our understanding of the world.

Let's imagine an old rustic wooden table. What is artificial about it? That which we can grasp with words: shape and size of its geometrical idealisation, its weight, colour tone, purpose, or perhaps a description of the way it was made by a carpenter with an axe, a saw and a jack plane. However, we cannot describe how *exactly* it looks, how it feels when being touched, the exact look of its texture and wood structure, its smell.

Now let's imagine a three-legged white round plastic garden table. How to grasp it with words? Just take its designer's drawings and the description of technological aspects of its manufacturing and we have it right in front of us. We do not need to see and touch and feel this table to fully know *how* and *what* it really is—hence it is almost completely artificial. Yet even such an artificial thing has something natural about it: various scratches, defects, imperfections, shabbiness, but most importantly its inherent qualia potential that we exploit when we meet the table right here and now. All these aspects defy being captured by words, and therefore are natural.

Through language, we can build scaffolding around the world. We build it step by step, further and further. We know that if we build a floor of the scaffolding, we can add one more. Yet we know that we can never reach the

sky; we can never breach the horizon—it would always become the chasing of a rainbow. But—at least we know everything about this scaffolding. We know everything about the world it encompasses, as much as we can know about a landscape from the map: it is not for feasting one's eyes on the beatiful countryside, but for perfect orientation it is quite enough. The scaffolding itself is very much artificial and can be exemplified as a particular domain of a scientific discourse. Those things in the scaffolded world, for which "feasting one's eyes" equals "perfect orientation", are purely artificial. The rest is still more or less pertaining to *physis*—especially the world beyond the horizon where the scaffolding does not reach.

However, what if we insist on building the scaffolding even beyond the horizon? We can construct a machine that will do it for us. The machine will pile up the scaffolding floors on top of each other so quickly that it will soon reaches the sky and even further. Or instead of the machine, we ourselves can put many big prefabricated scaffoldings on top of each other, hence going not step by step but by big leaps. This would also build the scaffolding beyond the horizon. But what is such a new scaffolding for us? We still stand where we were before and we know that we will never be able to climb up to the top to see how it looks beyond the horizon. The scaffolding itself thus ceases to be lucid for us anymore and starts to defy being captured by a (meta-)language. *Physis* strikes back. *Physis* again finds its way to the part of the world from which it was expelled.

In other words, when complexity of artificially built systems reaches a level on which it becomes impossible to describe them in finite time—to capture them by language—then the wild and chaotic world takes back what belongs to it anyway and those systems start to become natural. Maybe not at once, but naturalness gradually starts to proliferate through them.

This is exactly the trick of emergentism and emergent phenomena. All we need is *quantity*. Quantity beyond the horizon. A system may consist of purely artificial, perfectly describable, human-made elements. One such an element can be captured by language. Two of them as well. Three, four, five, ... still can be captured by language, hence still artificial. However, if the system consists of 100 billion such mutually interacting elements, it definitely cannot be captured by language—perhaps it can be captured by that superhigh scaffolding, but such a scaffolding cannot be captured itself, so it makes no difference. It is just like in sorites, "little-by-little" paradoxes—only there is nothing paradoxical about it; it is simply the phenomenological givenness of how we perceive the world. *Physis* thus comes back to the system, no matter the artificial in its elements. To put it simply: emergent phenomena are natural, not artificial.

If Artificial Intelligence (now we mean it as a "scientific discipline") creates an "artificial" mind emerging on top of an immensely complex system, this mind will be natural! As natural as our minds are. However, it will not be the AI engineers who are the authors or creators of its naturalness, who shall take the credit for it. The naturalness will be given to it from the same source and by the same means as it is given to everything else in the world. The AI engineers only prepare a substrate for it and then try to build the scaffolding high enough to lure the emergence through it.

AI research and development is metaphorically a Kabbalistic practice of its kind. A group of more or less wise men mould very complex inanimate matter, following strong rules, rituals and traditions, and then they ritually dance around this matter and heap up myriads of words arranged into very sophisticated spells, hoping that these words will evoke the spirit of emergence which brings naturalness and life into the artificial and inanimate.

This is the reason why GOFAI—Good Old-Fashioned Artificial Intelligence, i.e. "classical" AI in its symbolic, top-down paradigm—has not achieved to create anything natural. In GOFAI, the AI engineer is also The Creator, the one who knows how the system works and what it is that makes it intelligent, thinking, with mind. Therefore, the whole system is in front of the horizon, fully within the lucid structure of the scaffolding built by the engineer, fully captured by language—hence fully artificial. A man can be a creator, but only of the artificial.

Emergent AI is in a very different situation: naturalness leaks into artificially created systems through their immense complexity that lies far beyond the horizon of what can be captured by language. However, the AI engineer has a fundamentally different role here: he is not The Creator anymore, and he remains only a priest, sage, shamman, theurgist. He knows what he did but he does not know what exactly it is that makes the system intelligent, aware, sentient, thinking.

So what are our Artificial Intelligence dreams about? If they are about us being The Creators of new *natural* artificial intelligence and minds, then we really dream Artificial Dreams. Yet it is natural to dream Artificial Dreams, and perhaps even pleasant, comforting and helpful. But when we wake up from the dreams, we should seriously start to think how to live with the natural machine intelligence that has already started to emerge on top of our technological artifacts.

Beyond AI: Artificial Dreams
Proceedings of the International Conference Beyond AI 2012

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

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

Published by University of West Bohemia, Pilsen, 2012
First edition

ISBN 978-80-261-0102-4