Topics in Intelligent Engineering and Informatics 9

Jan Romportl
Eva Zackova
Jozef Kelemen *Editors*

Beyond Artificial Intelligence

The Disappearing Human-Machine Divide

Springer

Topics in Intelligent Engineering and Informatics

Volume 9

More information about this series at http://www.springer.com/series/10188

Jan Romportl · Eva Zackova
Jozef Kelemen
Editors

Beyond Artificial Intelligence

The Disappearing Human-Machine Divide

Springer

Editors

Jan Romportl
New Technologies Research Centre
& Department of Cybernetics
University of West Bohemia
Plzen
Czech Republic

Jozef Kelemen
Institute of Computer Science
Silesian University
Opava
Czech Republic

Eva Zackova
New Technologies Research Centre
University of West Bohemia
Plzen
Czech Republic

ISSN 2193-9411 ISSN 2193-942X (electronic)
ISBN 978-3-319-09667-4 ISBN 978-3-319-09668-1 (eBook)
DOI 10.1007/978-3-319-09668-1

Library of Congress Control Number: 2014945745

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This book is an edited collection of chapters based on the papers presented at the conference "Beyond AI: Artificial Dreams" held in Pilsen in November 2012 as the second in the three-year series of the *Beyond AI* conferences. The aim of the conference was to question deep-rooted ideas of artificial intelligence and cast critical reflection on methods standing at its foundations. Artificial Dreams – also an allusion on another book of one of our co-authors, Hamid Ekbia – epitomise our controversial quest for non-biological intelligence, and therefore the contributors of this volume tried to fully exploit such a controversy in their respective chapters, which resulted in an interdisciplinary dialogue between experts from engineering, natural sciences and humanities.

While pursuing the Artificial Dreams, it has become clear that it is still more and more difficult to draw a clear divide between human and machine. And therefore this book tries to portrait such an image of what lies beyond artificial intelligence: we can see the disappearing human-machine divide, a very important phenomenon of nowadays technological society, the phenomenon which is often uncritically praised, or hypocritically condemned. And so this phenomenon found its place in the subtitle of the whole volume as well as in the title of the chapter of Kevin Warwick, one of the keynote speakers at "Beyond AI: Artificial Dreams".

I would like to thank all who made the conference and the book happen. Special thanks go to all the authors, to my co-editors Eva Zackova and Jozef Kelemen, and to Pavel Ircing who tremendously helped all of us with the TEXnical issues.

Jan Romportl

Table of Contents

The Disappearing Human-Machine Divide

Kevin Warwick

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

Abstract. In this article a look is taken at three areas in which the divide between humans and machines is rapidly diminishing. A look is taken firstly at culturing biological neurons and embodying them within a robot body, secondly the use of implants to link a human nervous system with the Internet and thirdly recent results from the Turing Imitation Game which concentrates on differences in human communication. In each case the technical background is described, practical results are discussed and finally implications and future directions are considered.

Keywords: cyborgs, implant technology, bio-tech hybrids, human enhancement, Turing test.

1 Introduction

It is clear that as technology improves and human dependence on that technology increases so the gap between humans and machines is rapidly diminishing. To this end a focus of attention has been placed on interfaces between technology and the human brain. This is done from a practical perspective with applications in mind, however some of the implications are also considered. Results from experiments are considered in terms of their meaning and application possibilities. The article is written from the perspective of scientific experimentation opening up realistic possibilities to be faced in the future rather than giving conclusive comments. Human implantation and the merger of biology and technology are important elements, in the next two sections at least.

In this article different experiments in linking biology and technology together in a cybernetic fashion, essentially ultimately combining humans and machines in a relatively permanent merger, are considered. However a look is also taken, by means of the Turing test, at conversational abilities and how easy or difficult it is to tell the difference between humans and machines. Each of the sections involves practical experiments as something that have been actually realised, i.e. we are looking here at actual real world experiments as opposed to mere philosophical speculations.

The different experiment experiments are described in their own section. Whilst there is distinct overlap between the sections, they each throw up individual considerations. Following a description of each investigation some pertinent issues on the topic are therefore discussed.

J. Romportl et al. (eds.), *Beyond Artificial Intelligence,*
Topics in Intelligent Engineering and Informatics 9, DOI: 10.1007/978-3-319-09668-1_1

2 Biological Brains in a Robot Body

The first area considered might not be at all familiar to the reader. When one thinks of linking a brain with technology then it is probably in terms of a brain already functioning within its own body. Here however we consider the possibility of a fresh merger where a brain, consisting of biological neurons, is grown and then given its own body in which to operate.

An experimental control platform, a robot body, can move around in a defined area purely under the control of such a network and the effects of the brain, controlling the body, can be witnessed. Investigations can thus be performed into memory formation and reward/punishment scenarios – elements that underpin the functioning and growth mechanisms of a brain.

Growing networks of brain cells (around 100,000) in vitro begins by using enzymes to separate neurons obtained from foetal rodent cortical tissue. They are then grown (cultured) in a specialised chamber, in which they can be provided with controlled environmental conditions (e.g. appropriate temperature) and nutrients [1, 2]. An array of electrodes embedded in the base of the chamber (a Multi Electrode Array; MEA – see Figure 2) acts as a bi-directional electrical interface with which to provide signals to the culture and to monitor signals from the culture. This enables electrical signals to be supplied both for input stimulation and also for recordings to be taken as outputs from the culture. The neurons in such cultures spontaneously connect, communicate and develop, within a few weeks.

With the MEA it is possible to separate the firings of small groups of neurons by monitoring the output signals on the electrodes. Thereby a picture of the global activity of the brain network can be formed. It is also possible to electrically stimulate the culture via any of the electrodes to induce neural activity. The multi-electrode array therefore forms a bi-directional interface with the cultured neurons [3, 4].

The brain is coupled to its physical robot body [5]. Sensory data fed back from the robot is delivered to the culture, thereby closing the robot-culture loop. The processing of signals can be broken down into two discrete sections (a) ‘culture to robot’, in which live neuronal activity is used as the decision making mechanism for robot control, and (b) ‘robot to culture’, which involves an input mapping process, from robot sensor to stimulate the culture.

The number of neurons in a brain depends on natural density variations in seeding the culture in the first place. The electrochemical activity of the culture is sampled and is used as input to the robot’s wheels. The robot’s (ultrasonic) sensor readings are converted into stimulation signals received by the culture, closing the feedback loop.

Once the brain has grown for several days, an existing neuronal pathway through the culture is identified by searching for strong relationships between (input-output) pairs of electrodes. A rough input-output response map of the culture can be created by cycling through the electrodes in turn. In this way, a suitable input/output electrode pair can be chosen in order to provide an initial decision making pathway for the robot. This is then employed to control

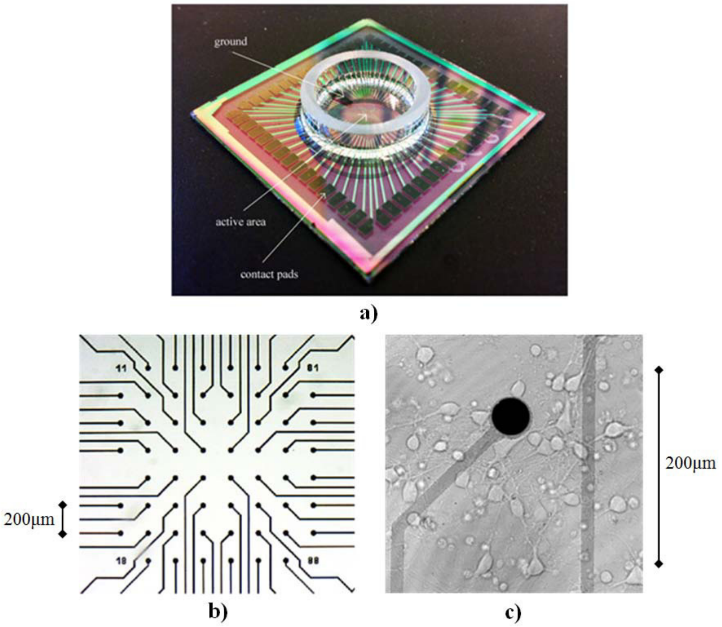

Fig. 1. a) A Multi Electrode Array (MEA) showing the electrodes b) Electrodes in the centre of the MEA seen under an optical microscope c) An MEA at x40 magnification, showing neuronal cells in close proximity to an electrode

the robot body – for example if the ultrasonic sensor is active and we wish the response to cause the robot to turn away from the object being located ultrasonically (possibly a wall) in order to keep moving.

For experimentation purposes at this time, the robot must follow a forward path until it nears a wall, at which point the front sonar value decreases below a threshold, triggering a stimulating pulse. If the responding/output electrode registers activity the robot turns to avoid the wall. The most relevant result is the occurrence of the chain of events: wall detection-stimulation-response. However from a neurological perspective it is of also interesting to speculate why there is activity on the response electrode when no stimulating pulse has been applied.

The cultured brain acts as the sole decision making entity within the overall feedback loop. Clearly one important aspect then involves neural pathway changes, with respect to time, in the culture between the stimulating and recording electrodes. Learning and memory investigations are generally at an early stage. However the robot can be witnessed to improve its performance over time in terms of its wall avoidance ability in the sense that neural pathways that bring about a satisfactory action tend to strengthen purely though the process of being habitually performed – learning due to habit.

The number of variables involved is considerable and the plasticity process, which occurs over quite a period of time, is dependent on such factors as initial seeding and growth near electrodes as well as environmental transients such as temperature and humidity. Learning by reinforcement – rewarding good actions and punishing bad is merely investigative research at this time.

It has been shown by this research that a robot can successfully have a biological brain with which to make its 'decisions'. The culture size is merely due to the present day limitations of the experimentation described. Indeed 3 dimensional structures are presently being investigated. Increasing the complexity from 2 dimensions to 3 dimensions realises a figure of over 30 million neurons for the 3 dimensional case – not yet reaching the 100 billion neurons of a perfect human brain, but well in tune with the brain size of many other animals.

Not only is the number of cultured neurons increasing, but the range of sensory input is being expanded to include audio and visual. Such richness of stimulation will no doubt have a dramatic effect on culture development. The potential of such systems, including the range of tasks they can deal with, also means that the physical body can take on different forms. There is no reason, for example, that the body could not be a two legged walking robot, with rotating head and the ability to walk around.

At present rat neurons are usually employed in studies. However human neurons are also now being cultured, allowing for the possibility of a robot with a human neuron brain. If this brain then consists of billions of neurons, many social and ethical questions will need to be asked [6]. For example – if the robot brain has roughly the same number of human neurons as a typical human brain then could/should it have similar rights to humans? Also – what if such creatures had far more human neurons than in a typical human brain – e.g. a million times more – would they make all future decisions rather than regular humans?

3 Braingate Implant

It is perhaps often the case that brain-computer interfaces are used for therapeutic purposes, to overcome a medical/neurological problem. However there is also the possibility to employ such technology to give individuals abilities not normally possessed by humans (cf. Chapter 3 in this volume).

Some of the most impressive human research to date in this area has been carried out using the microelectrode array, shown in Figure 2. The individual electrodes are 1.5 mm long and taper to a tip diameter of less than 90 microns. Although a number of trials not using humans as a test subject have occurred, human tests are at present limited to a small group of studies. In some of these the array has been employed in a recording only role, most notably as part of (what was then called) the 'Braingate' system.

Electrical activity from a few neurons monitored by the array electrodes was decoded into a signal to direct cursor movement. This enabled an individual to position a cursor on a computer screen, using neural signals for control combined with visual feedback. The same technique was later employed to allow the individual recipient, who was paralysed, to operate a robot arm [7, 8].

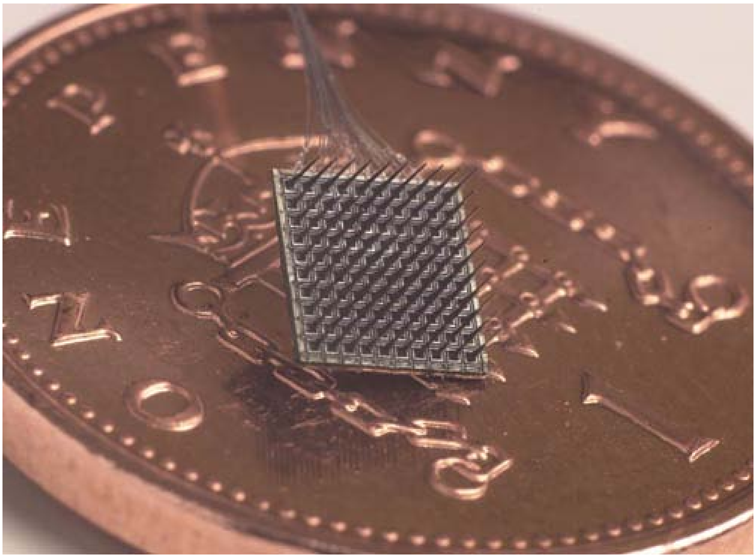

Fig. 2. A 100 electrode, 4X4mm Microelectrode Array, shown on a UK 1 pence piece for scale

The first use of the Braingate microelectrode array (shown in Figure 2) in a human has though considerably broader implications which extend the capabilities of the human recipient. The array was implanted into the median nerve fibres of a healthy human individual (the author) during two hours of neurosurgery in order to test bidirectional functionality in a series of experiments. A stimulation current directly into the nervous system allowed information to be received, while control signals were decoded from neural activity in the region of the electrodes [9, 10]. A number of experimental trials were successfully concluded [11, 12], in particular:

1. Extra sensory (ultrasonic) input was successfully implemented.
2. Extended control of a robotic hand across the Internet was achieved, with feedback from the robotic fingertips being sent back as neural stimulation to give a sense of force being applied to an object (this was achieved between Columbia University, New York (USA) and Reading University, England).
3. A primitive form of telegraphic communication directly between the nervous systems of two humans (the author's wife assisted) was performed [12].
4. A wheelchair was successfully driven around by means of neural signals.
5. The colour of jewellery was changed as a result of neural signals – also the behaviour of a collection of small robots.

In all of the above cases it can be regarded that the trial proved useful for purely therapeutic reasons, e.g. the ultrasonic sense could be useful for an individual who is blind or the telegraphic communication could be useful for those with certain forms of Motor Neurone Disease. However each trial can also be seen as a potential form of enhancement beyond the human norm for an individual. The author did not need to have the implant for medical purposes to overcome a problem but rather for scientific exploration.

From the trials it is clear that extra sensory input is one practical possibility that has been successfully attempted, however improving memory, thinking in many dimensions and communication by thought alone are other distinct potential, yet realistic, benefits, with the latter of these also having been investigated to an extent. To be clear – all these things appear to be possible (from a technical viewpoint at least) for humans in general.

An individual human connected in this way can potentially also benefit from some of the advantages of machine/artificial intelligence, for example rapid and highly accurate mathematical abilities in terms of 'number crunching', a high speed, almost infinite, Internet knowledge base, and accurate long term memory. Humans are also limited in that presently they can only visualise and understand the world around them in terms of a limited 3-dimensional perception, whereas computers are quite capable of dealing with hundreds of dimensions.

Most importantly, the human means of communication, essentially transferring a complex electro-chemical signal from one brain to another via an intermediate, often mechanical slow and error prone medium (e.g. speech), is extremely poor, particularly in terms of speed, power and precision. It is clear that connecting a human brain, by means of an implant, with a computer network could in the long term open up the distinct advantages of machine intelligence, communication and sensing abilities to the implanted individual.

4 Turing Imitation Game

The final area to be considered is that of practical Turing tests which give an indication of how easy or difficult it is to distinguish between humans and machines in terms of conversational ability. In this article I have concentrated on the comparison test as originally described by Turing in describing his Imitation Game [13]. It is worth remembering that Turing originally proposed the test as a replacement for the question "Can Machines Think?" [13], however here I am more concerned with the practical nature of the test rather than in any philosophical argument with regard to its meaning.

The test was described by Turing himself in 1952 as: "The idea of the test is that a machine has to try and pretend to be a man, by answering questions put to it, and it will only pass if the pretence is reasonably convincing. A considerable portion of a jury, who should not be expert about machines, must be taken in by the pretence" [14, p. 495].

The Turing test involves a machine which pretends to be a human in terms of conversational abilities. In a paired comparison the attempt is for the machine to appear to be more human than the human against whom it is paired. To conform to Turing's original wording in his 1950 paper [13] I refer here to 5 minute long tests only. I am well aware that there are those who take issue over a suitable timing and what Turing actually meant [15] – that is considered to be an argument for another day, it does not alter the point made here.

What is presented here are three specific transcripts selected from a day of actual, practical Turing tests which were held under strictly timed conditions with many external viewers at Bletchley Park, England on 23rd June 2012. The date marked the 100th anniversary of Turing's birth and the venue was that at which amongst other things, during the Second World War, Turing led a team of codebreakers who cracked the German Enigma machine cypher. Five different machines took part in the tests during the day along with 30 different judges and numerous hidden humans against which the machines were compared in terms of their conversational ability.

What I focus on here is not how good or bad the machines were at deception or how human the hidden humans were but rather the decisions taken by the judges and how these might compare with your own selections. So this article is more a look at the differences between the hidden entities and how easy or difficult it can be to tell which is human and which is machine.

What follows are three separate transcripts. These represent actual transcripts taken on the morning of 23rd June 2012 at Bletchley Park, England between different human judges/interrogators and hidden entities. Each conversation lasted for a total of 5 minutes exactly and no more, just as Turing stipulated [13]. There was a hard cut off at that time and no partial sentences were transmitted. Once a sentence had been transmitted it could not be altered or retracted in any way. Hence all wording and spelling is exactly as it was at the time – any spelling mistakes are those which actually occurred in the test, they are not due to poor editing.

Judges were asked to decide if the entity was human, machine or if they were unsure. The hidden humans meanwhile were asked to be themselves – human – they were asked not to make it easy for the machines because it was the machines which were competing against them for humanness [16]. They were however asked not to reveal personal details about themselves that could in any way give away their actual identity to a judge.

What I show below are three specific selected transcripts on their own, i.e. not in parallel paired fashion [17]. The reason for this is that the aim here is not to make a direct comparison between these transcripts and any others, but rather to analyse these transcripts in their own right in order to see if you can judge correctly as to the identity of the entity.

At the end of this section the judges conclusions on each transcript are given. These were the conclusions actually drawn by those judges. They do not reflect my own conclusions on each of the transcripts. Following this the true identity of each entity is given. In each case the transcript involved a different judge.

Transcript 1

Judge: Hi, here we go. How do you feel about this encounter?
Entity: i can only comment once we've had sufficient of it
Judge: Fair enough! Anyway, I almost didn't get here today.

Entity: and why is that?
Judge: I felt very dissy when I woke up and it was hard to shake off.
Entity: and you had to make it on time anyways. good job!
Judge: thanks. by the way I meant dizzy. Bad typing. I had to drive faster than normal.
Entity: and you didn't get any speed limit fine
Judge: no, luckily. There was an interesting thing on the Today programme this morning.
Entity: and what is it? this encounter?
Judge: you mean the thing on the programme, or the chat we're having?

Transcript 2 (taken from [18])

Judge: hey, just to lighten the mood a little, what's brown and sticky?
Entity: no idea, what's the answer?
Judge: a stick... so anyway what's your name?
Entity: Sheldon Cooper, what's yours?
Judge: as in the guy off the big bang theory?
Entity: I don't believe in the big bang
Judge: so are you religious then?
Entity: aren't we all?
Judge: well i like to consider myself an agnostic, which religion is it you follow then if you dont mind me asking?
Entity: you can ask what you like, it's a free world, who am I to mind?

Transcript 3

Judge: Hello. How are you?
Entity: Hey.
Judge: How do you like Bletchley Park?
Entity: lol.
Judge: Are you from England?
Entity: They have Wi-Fi here in the pub.
Judge: Which pub?
Entity: I'm just down the pub.
Judge: Have you ever been in a Turing Test before?
Entity: Will this be the 5 minute argument, or were you thinking of going for the full half hour.
Judge: Very funny. You sound suspiciously human. Do you like the Beatles?
Entity: I'd like to get the next Dread the Fear tape.
Judge: What is Dread the Fear?
Entity: Dread the fear has that Steve Henderson guy in it.
Judge: What sort of music is that? Or is it comedy?

Solutions
In transcript 1 the judge was a male expert in Artificial Intelligence who was very familiar with such tests. Here he decided that the hidden entity was a machine. In fact the entity was a male human. In transcript 2, with a different judge,

again the decision was that the entity was a machine whereas in reality it was a human – the author of this article in fact. In transcript 3 meanwhile the judge's decision was that the hidden entity was a male human whereas it was in fact a machine. So how did you do?

5 Conclusions

In this article a look has been taken at different ways in which the human-machine divide is diminishing. It has to be said that these are all perhaps a curious selection as examples and many other instances exist, particularly so in everyday life. Rather than focus merely on theory, here practical experimental cases have been reported on. Further details for each of these can be found in a variety of publications. In each case questions arise as a result.

When considering robots with biological brains, this could ultimately mean human brains operating in a robot body. Therefore, should such a robot be given rights of some kind? If one was switched off would this be deemed as cruelty to robots? More importantly at this time – should such research forge ahead regardless? Before too long we may well have robots with brains made up of human neurons that have the same sort of capabilities as those of the human brain – is this acceptable?

In the section focusing on the Braingate implant as a general purpose invasive brain implant, as well as its employment for therapy a look was taken at the potential for human enhancement. Already extra-sensory input has been scientifically achieved, extending the nervous system over the Internet and a basic form of thought communication. So if many humans upgrade and become part machine (Cyborgs) themselves, what would be wrong with that? If ordinary (non-implanted) humans are left behind as a result then what is the problem? If you could be enhanced, would you have any problem with it?

In the final section some of the latest results from the Turing test were presented. In each of the three cases mentioned so the judge drew an incorrect conclusion. Even if you did manage to get three correct answers out of three hopefully you are able to agree, with these transcripts as examples, that machine conversation is now getting to the stage where it is difficult for an external observer to decide which is human and which is machine. It has to be said though that this is just as much down to the fallibility of humans as it is to the present-day wonders of machine communication.

References

1. Chiappalone, M., Vato, A., Berdondini, L., Koudelka-Hep, M., Martinoia, S.: Network dynamics and synchronous activity in cultured cortical neurons. Int. J. Neural Syst. 17(2), 87–103 (2007)
2. De Marse, T.B., Wagenaar, D.A., Blau, A.W., Potter, S.M.: The neurally controlled animat: Biological brains acting with simulated bodies. Autonomous Robots 11, 305–310 (2001)

3. Warwick, K., Nasuto, S.J., Becerra, V.M., Whalley, B.J.: Experiments with an in-vitro robot brain. In: Cai, Y. (ed.) Computing with Instinct 2010. LNCS (LNAI), vol. 5897, pp. 1–15. Springer, Heidelberg (2011)
4. Warwick, K., Xydas, D., Nasuto, S.J., Becerra, V.M., Hammond, M.W., Downes, J., Marshall, S., Whalley, B.J.: Controlling a mobile robot with a biological brain. Defence Science Journal 60(1), 5–14 (2010)
5. Xydas, D., Norcott, D.J., Warwick, K., Whalley, B.J., Nasuto, S.J., Becerra, V.M., Hammond, M.W., Downes, J., Marshall, S.: Architecture for neuronal cell control of a mobile robot. In: European Robotics Symposium, EUROS 2008, pp. 23–31 (2008)
6. Warwick, K.: Implications and consequences of robots with biological brains. Ethics and Information Technology 12(3), 223–234 (2010)
7. Donoghue, J.P., Nurmikko, A., Friehs, G., Black, M.: Development of neuromotor prostheses for humans. Supplements to Clinical Neurophysiology 57, 592–606 (2004)
8. Hochberg, L.R., Serruya, M.D., Friehs, G.M., Mukand, J.A., Saleh, M., Caplan, A.H., Branner, A., Chen, D., Penn, R.D., Donoghue, J.P.: Neuronal ensemble control of prosthetic devices by a human with tetraplegia. Nature 442(7099), 164–171 (2006)
9. Warwick, K., Gasson, M.N., Hutt, B., Goodhew, I., Kyberd, P., Andrews, B., Teddy, P., Shad, A.: The application of implant technology for cybernetic systems. Archives of Neurology 60(10), 1369–1373 (2003)
10. Gasson, M.N., Hutt, B.D., Goodhew, I., Kyberd, P., Warwick, K.: Invasive neural prosthesis for neural signal detection and nerve stimulation. International Journal of Adaptive Control and Signal Processing 19(5), 365–375 (2005)
11. Warwick, K., Gasson, M.: Practical interface experiments with implant technology. In: Sebe, N., Lew, M., Huang, T.S. (eds.) ECCV/HCI 2004. LNCS, vol. 3058, pp. 7–16. Springer, Heidelberg (2004)
12. Warwick, K., Gasson, M.N., Hutt, B., Goodhew, I., Kyberd, P., Schulzrinne, H., Wu, X.: Thought communication and control: A first step using radiotelegraphy. IEE Proceedings - Communications 151(3), 185–189 (2004)
13. Turing, A.M.: Computing machinery and intelligence. Mind LIX(236), 433–460 (1950)
14. Copeland, B.J.: The Essential Turing: The Ideas that Gave Birth to the Computer Age. Oxford University Press (2004)
15. Shah, H., Warwick, K.: Testing Turing's five minutes, parallel-paired imitation game. Kybernetes 39(3), 449–465 (2010)
16. Warwick, K.: Not another look at the Turing test? In: Bieliková, M., Friedrich, G., Gottlob, G., Katzenbeisser, S., Turán, G. (eds.) SOFSEM 2012. LNCS, vol. 7147, pp. 130–140. Springer, Heidelberg (2012)
17. Shah, H., Warwick, K.: Hidden interlocutor misidentification in practical Turing tests. Minds and Machines 20(3), 441–454 (2010)
18. Warwick, K., Shah, H., Moor, J.: Some implications of a sample of practical Turing tests. Minds and Machines, 1–15 (2013)

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, 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: artificial intelligence, 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

* 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.

J. Romportl et al. (eds.), *Beyond Artificial Intelligence,*
Topics in Intelligent Engineering and Informatics 9, DOI: 10.1007/978-3-319-09668-1_2

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 expert that experts would *not* be showing particular skill 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 thought 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 be 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[1][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 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

[1] 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.

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.[2] 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

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

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 timeline of previous revolutions [24]? If we can't extract assumptions, we're reduced to saying "that feels 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, an extra hidden assumption"). The counterfactual resiliency exercise has been carried out at length in an online post.[3] 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 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

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

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 gives 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, programming 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 modelling grind should be accepted more than predictions that are modelling insight? Not at all. Predictions that are modelling 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 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 25AI 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[4]). 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 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 foresaw 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

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

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.[5]

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.)

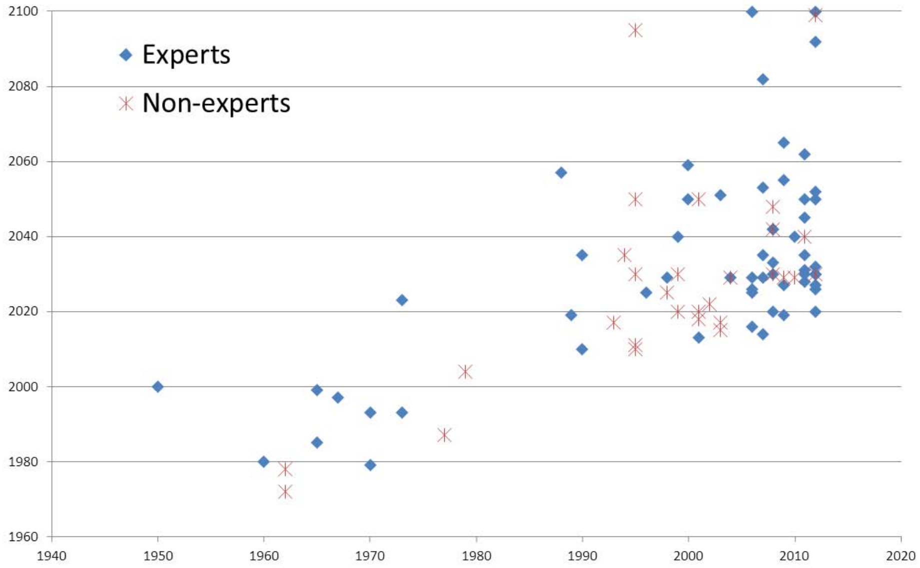

Fig. 1. Median estimate for human-level AI, graphed against date of prediction

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.

[5] 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.

Two explanations have been generally advanced to explain poor expert performance in these matters. The first, the so-called Maes-Garreau law[6] 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 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).

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.

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.

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

[6] 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).

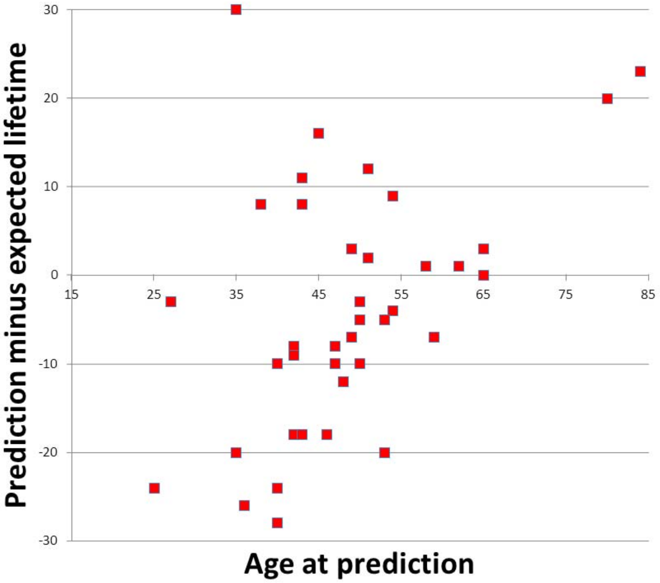

Fig. 2. Difference between the predicted time to AI and the predictor's life expectancy, graphed against the predictor's age

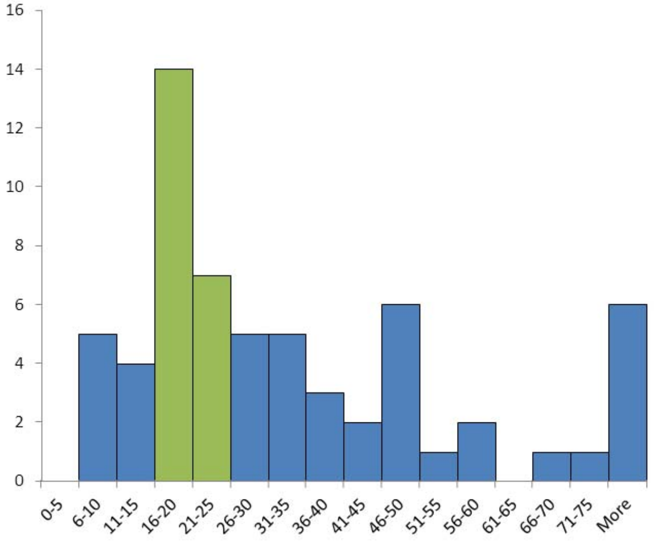

Fig. 3. Time between the arrival of AI and the date the prediction was made, for expert predictors

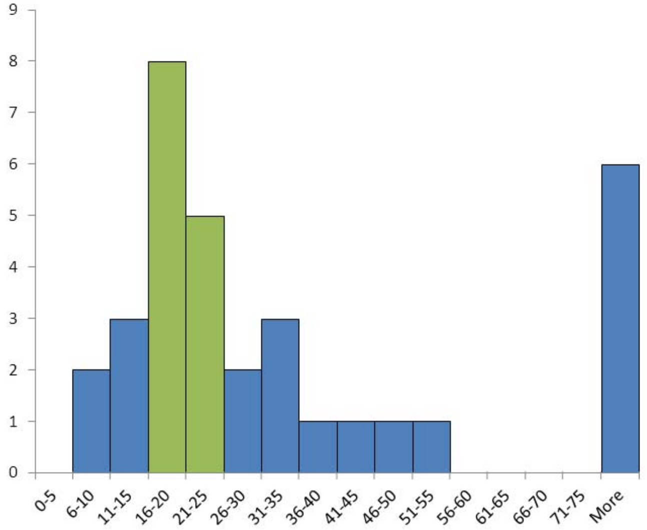

Fig. 4. Time between the arrival of AI and the date the prediction was made, for non-expert predictors

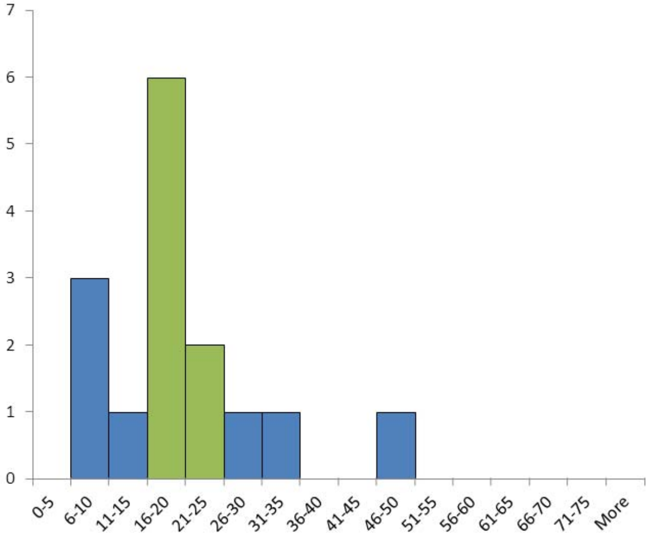

Fig. 5. Time between the arrival of AI and the date the prediction was made, for failed predictions

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.

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.S.: Further reflections on the timescale of AI. In: Dowe, D.L. (ed.) Solomonoff Festschrift. LNCS, vol. 7070, pp. 174–183. Springer, Heidelberg (2013)
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, vol. 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.: 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.: Über 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)

Intelligence Explosion Quest for Humankind*

Eva Zackova

Department of Interdisciplinary Activities, New Technologies Research Centre
University of West Bohemia, Pilsen, Czech Republic
zacka@ntc.zcu.cz

Abstract. Traditionally, the discipline of artificial intelligence (AI) aims to creation of artificial intelligent entity that will be (at least) adequate (if not superior) to human intellectual power. In spite of this fact, we witness progress heading in quite a different direction, towards fusion of human and specialized AI-based systems. Vernon Vinge called this process as intelligence amplification (IA) which he regarded as an alternative way of how to achieve a greater-than-human intelligence without an existential risk for humankind possibly coming from AI. In this paper, we advocate this scenario by propounding an overview of deep-rooted conceptions of human cyborgization. Such technological enhancement is interpreted as profoundly supportive of intelligence amplification conception. In comparison to AI, IA is regarded as more probable and desirable future for the mankind.

Keywords: intelligence explosion, strong artificial intelligence, weak artificial intelligence, intelligence amplification, artificial general intelligence, cyborg, man-computer symbiosis, singularity, transhumanism, human enhancement.

1 Introduction

Traditionally, the discipline of artificial intelligence (AI) aims at creation of artificial intelligent entity that will be (at least) adequate (if not superior) to human intellectual power. Despite this fact, we witness progress heading in quite a different direction, towards fusion of human and specialized AI-based systems. In the broad sense, this fusion started in the moment of the mankind rising and has been developing through technogenetic spiral since. Nevertheless, today we participate on unprecedented invasive proliferation of AI-based technologies into our brains and bodies. Generally speaking, human cognitive capabilities, intellect and physical power are being commonly technologically improved and amplified.

In comparison to various artificial intelligence conceptions, the idea of intelligence amplification (IA) is more and more frequently regarded as more plausible,

* This work was supported by 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.

J. Romportl et al. (eds.), *Beyond Artificial Intelligence,*
Topics in Intelligent Engineering and Informatics 9, DOI: 10.1007/978-3-319-09668-1_3

more realistic and even safer way of application and development of the AI field. As we progress further in AI, we can hear stronger and stronger voices alerting us to an existential risk for humankind rising from the development of artificial (general) intelligence (AGI) that might end up in a so-called technological singularity. The biggest advantage of IA dwells in its potential to lower such risk and even to avoid it completely. We simply have to deal with coming of an enormous intelligence explosion, and IA seems to offer a solution of this possibly dangerous epitomization of Moore's law. Following the genesis of technology that is getting closer to truly intimate connection with man, we create a *superman.* The forthcoming posthuman age of cyber-humanity is no longer a sci-fi movie. It is the most probable future of our own species.

At the present time, contemplation on the dreamed-of universal general artificial intelligence is more often left behind either as non-feasible or as a deadly dangerous goal, whereas technological enhancement of human intellect seems to be inevitable, and moreover, it already occupies appreciable part of the AI field.

2 Machina Sapiens

One of the very first definitions of artificial intelligence was made by John McCarthy in 1956:

> The study [of AI] is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. [1, p. 17]

The core idea of this definition actually referring to a research project that was supposed not to exceed two-month work of ten-member team, is valid in the AI field until these days:

> We call ourselves *Homo sapiens* – man the wise – because our **intelligence** is so important to us. For thousands of years, we have tried to understand *how we think*; that is, how a mere handful of matter can perceive, understand, predict, and manipulate a world far larger and more complicated than itself. The field of **artificial intelligence**, or AI, goes further still: it attempts not just to understand but also to *build* intelligent entities. [1, p. 1]

Very soon after the birth of the AI field, its original goal, i.e. to simulate intelligence and rationality of humans, divided into various specific tasks that we can classify into several categories such as learning and perception, knowledge representation, communication and agency, automatic speech processing, computer vision, pattern recognition, real-time interaction, orientation tasks and so on. Even more, such AI systems focused on particular tasks are frequently domain restricted (health and medicine area, bus departures, chess playing, furniture ordering, weather conversation and so on). Despite the fact that AI engineers mostly focus on such a particular and specific problem solving tasks, still we can regard their effort as related to the original bigger intention to understand principles of human mind functioning, and based on this, to simulate the mind in a nonbiological substrate of a computer.

Later, this goal was emphasized again in the conception of the so-called human-level AI [1, p. 27] (which is discussed in relation to a prediction in Chapter 2 of this volume). Even more popular is to discuss the related conception of *artificial general intelligence* (AGI) currently proposed mainly by Ben Goertzel and Ray Kurzweil. Russell and Norvig define AGI as a field, *which looks for a universal algorithm for learning and acting in any environment* [1, p. 27]. Briefly, AGI is not just about putting parts together, it aims to creation of truly universal self-learning machine.

In 1966 John Good considered possible results of creating AGI and came to a conclusion that such AGI would be very probably endowed by the power of self-improving ability which could lead to a continuous chain of self-refinement steps heading to the intelligence explosion:

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an ‘intelligence explosion’, and the intelligence of man would be left far behind ... Thus the first ultraintelligent machine is the *last* invention that man need ever made ... [2, p. 33]

According to Good [2, p. 33], an important part of this vision is a presumption of AI’s docility towards humans. This idea is very controversial and we will see that many thinkers expect an exact opposite state of affairs in the future, they warn us about AI going far beyond human dimension and control. *Singularitarianism* is the one focused on such issues. It is a futuristic school of thought dealing with the intelligence explosion scenario and its possible effects on human race and its future existence on Earth.

Philosophy of mind is another field maintaining the human mind, consciousness, rationality, intelligence, and artificial intelligence as well. From philosophical point of view, one of the most fundamental question is whether AI can *truly* think and be endowed with consciousness. When a philosopher talks about such a truly phenomenologically conscious (yet artificially created) mind, he uses the label *strong artificial intelligence*; whereas when he uses notion *weak artificial intelligence* it means that he regards the referred entity as a mere simulation of an intelligent behavior. The philosophical distinction between weak and strong AI was coined by John Searle. He introduced it in his famous *Chinese Room* thought experiment [3].

In practice, an engineer or computer scientist working on an AI system would be more than happy if such an AI just *simulated* human intelligence and its general universality at least a bit. He is not interested in the phenomenological dimension of AI nor issues of mind at all. He is focused on providing required outputs of the AI system. It is tempting to identify the human-level AI conception or AGI conception with the philosophical notion of strong AI, but in fact the theory of AGI does not say anything about consciousness or mind explicitly. It is hard to draw a line between these two (engineering and philosophical) conceptions because they overlap in many aspects but they are totally incomparable in the others.

The only thing that seems to be plausible enough is that the strong AI conception presumes the creation of AGI (because it is not easy to award a consciousness to just a chess-playing program indeed). We do not know how to measure amount of *psyché* in computers (AGI) nor if it is really necessary for performing universal intelligence. Actually, it is quite complicated to prove the consciousness even in case of human brains, as we know from discussions on *philosophical zombies*. The tendency to regard as granted that a general AI automatically means a strong AI is probably based on our anthropomorphizing attitude and applied intersubjectivity which represent our intuitive and of course understandable approach to AI – but still need a serious argumentation at the same time. Categorical analyses of notions such as human being, machine, rationality, intelligence, life, mind, consciousness etc. extend the range of interesting and important questions within epistemology, ethics, law and humanities in general which focus their effort on reflecting strong AI and its relation to human civilization.

Even though nobody knows right now how to make such self-conscious intelligence come to an existence, we can get a glimpse of what might be the only way of AI emergence in Jan Romportl's chapter *Naturalness in Artificial Intelligence* in this volume. We have not given up thinking and designing a genuine *machina sapiens* in the field of philosophy as well. Yet dreaming of AI, we can easily find ourselves in the middle of a night-mare scenario heading to an extinction of human race, or at least its humiliation. Most famously, this is the case of Kevin Warwick's *March of The Machines* [4], and others who emphasize the intellectual potential of AI to significantly exceed the human kind of intelligence far beyond its limits and our imagination (see [5, 6]). Thus, they give us more or less solid reasons for not just to dream but to discuss and consider consequences of such an AI entity coming to our world, with serious concerns and interest.

3 Machines and Natural Selection

We could start counting numerous thinkers from the past that conduced to the idea of AI. After all, we always try to find roots of big thoughts, even though we might know that our interpretation is often inappropriate, and serves just our pragmatic purpose which is to justify the thought through discovering purported tradition. In spite of my doubts about these elaborated histories, I want to bring out 150 years old Samuel Butler's unexpectable vision of future man and *intelligent* machine coexistence that I have found exceptionally far-sighted for that time.

The text we are talking about was published in 1863 and from the present point of view, it brims over with highly futuristic thoughts. In that time, in the half of the 19th century, we didn't have electric iron yet, and foundations of computer sciences have just glimpsed in Charles Babbage's and Ada Lovelace's hands.

As he acknowledged right in the title *Darwin Among the Machines* [5], Butler based his essay on Charles Darwin's theory of evolution that was published few years before (in 1859) in one of the most famous scientific works ever – *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* [7].

According to Butler, the answer to the question about future descendants of human species is simple. In comparison to animal and vegetable realm, he finds machines to be the most progressive species in history of the planet Earth. Butler expressed his expectations of further development of machines and humans clearly. In accordance with laws of natural selection, machines will become superior to us due to their extreme intelligence, steadfast moral qualities, high-level self-regulation and inner organization, self-acting power and freedom from emotional unbalance. Butler even identified size diminution as one of the very crucial aspects of technological progress, which is more than an obvious fact today. He called the machines *glorious creatures* or even *glorious animals* and suggested to create a classification of their whole kind in the manner of animal and plant taxonomies known at that time [5, pp. 182–183].

Despite the declared inferior role of humans as machines' servants in this scenario, Butler expected that conditions of human living would be much better in general thanks to the enlightened artificial minds who will *domesticate* humans and treat them as *we treat our horses, dogs, cattle, and sheep, on the whole, with great kindness ...* [5, p. 183]. Well, that assumption was naive in the face of modern industrial processing of meat and other animal products. Even though we know very well that the current state is not the ideal, this is what happens commonly in our world and may happen in the future as well, but if we accept Butler's point of view, the future of man does not appear to be so cruel as we are used to imagining today. Machines will rule over the world but humans will not become extinct. They will be needed for maintenance of the machines, and they will even assist their reproduction. Based on the reciprocal relationship between humans and machines, humans will enhance their species and improve their quality of life. They just will not be the rulers of the world. Instead, machinery will take command. As Butler pointed out, it is both paradoxical and natural in this situation, that *we are ourselves creating our own successors* [5, p. 182].

For a techno-optimistic mind, the very first reading of Samuel Butler's essay could be little tricky. It seems to be almost a utopian description of man-machine symbiosis, until one reads further and gets to Butler's quite radical, surprising and prudent resolution. He took the coming of super-intelligent machines for granted and thus he called for a prompt solution:

> Our opinion is that war to death should be instantly proclaimed against them. Every machine of every sort should be destroyed by the well-wisher of his species. Let there be no exceptions made, no quarter shown; let us at once go back to the primeval condition of the race. [5, p. 185]

If you think that it is not possible, no matter which side you have chosen in this war, then Butler has bad news for you:

> If it be urged that this is impossible under the present condition of human affairs, this at once proves that the mischief is already done, ... that we have raised a race of beings whom it is beyond our power to destroy, and that we are not only enslaved but absolutely acquiescent in our bondage. [5, p. 185]

I am sure that many would sign under these statements today, especially neo-luddites, but I will address rather those concepts of mutual influence and common development of AI and humankind which can be regarded as a more optimistic and even beneficious alternatives.

Many of these alternatives have found common ground in the opinion that humans, technologies and technics have been developing together and through each other from everlasting. Katherine Hayles uses the term *technogenetic spiral* to refer to this phenomenon [8]. Moreover, as Hayles emphasizes, for paleoanthropologist it is a commonplace fact that anthropogenesis and technogenesis goes hand in hand. Our further assertions in this paper are based on this presumption as well.

4 Singularitarianism

The notion of *singularity* related to AI is very well known today thanks to Ray Kurzweil [6], but the idea of continuously accelerating progress of technologies (especially AI) heading to a world inaccessible to our imagination was introduced earlier. Probably the biggest credit for spreading the word about so-called *technological singularity* takes Vernon Vinge who popularized the idea in his science fiction publications during the 1980's and later in his famous academic paper *The Coming Technological Singularity* [9].

In his paper, Vinge refers to John von Neumann and John Good who are believed to be the ones of the very first using the term singularity, even though not in a strictly physical or mathematical sense, but still with main emphasis on progress acceleration. In general, technological singularity is defined as *ever accelerating progress of technology and changes in the mode of human life* [9, p. 13].

Within all singularity-related discussions, we can usually identify two main approaches. None of them is optimistic in terms of future condition of the human species. The worst case scenario based on calculation of accelerating speed of technological progress and huge amount of resources needed for that kind of development, leads to a self-consumption of the whole planet Earth, humankind included. It will be an impalpable quick process and most probably, we will not even notice nor realize this is it.

The little less but still catastrophic scenario gives us some decent time to actually experience the coming Singularity, and thus it puts a question mark on our everyday life and relationship to those intelligent artificial entities. It will not be the same life as we know it now, that is taken for granted.

According to Vinge, the coming of Singularity means coming of *strongly superhuman intelligence* [9, p. 13], which should be understood not just as an improvement on quantitative level (such as speed or amount of processed data) in comparison to human intelligence, but as a fundamental qualitative change that brings greater than human intelligence in the sense of supra- or trans-.

Vinge addressed four possible ways of making the Singularity happen, and yes – he is sure that it will certainly come. One of the scenarios relies on biological means for improving our natural human intelligence. The remaining are dependent on development of computer hardware and progress in the field of AI, and we can split them into those two categories of the more and the less catastrophic Singularity futures mentioned few lines above. An instance of the worst case scenario would be a pure AI (the strong AI, using terms of philosophy), and shortly after that it is reasonable to expect arrival of AI+, then AI++ and so on (Ivan Havel elaborates on this with deep insight in [10]).

Another way described by Vinge that leads to the Singularity era of suprahuman intelligent beings, is actually a fusion of man and machine, and a transition to Post-human age.

> ... there are other paths to superhumanity. Computer networks and human-computer interfaces seem more mundane than AI, and yet they could lead to the Singularity. I call this contrasting approach Intelligence Amplification (IA). ... I am suggesting that we recognize that in network and interface research there is something as profound (and potentially wild) as Artificial Intelligence. With that insight, we may see projects that are not as directly applicable as conventional interface and network design work, but which serve to advance us toward the Singularity along the IA path. [9, p. 17]

Commenting on ethical and safety issues of IA, Vinge admits that we cannot foresee too much actually. Definitely, IA cannot be taken as a guaranteed path to a harmless future of humanity.

> The problem is not that Singularity represents simply the passing of humankind from central stage, but that it contradicts some of our most deeply held notions of being. [9, p. 19]

Actually, a significant change of many other notions, such as mind, intelligence, personal identity, privacy, corporeality or death and humanity, is needed in face of IA that has already started to realize and influence individual lives.

Recalling Alvin Toffler's *wave theory* describing three main stages of a society and its culture, we have probably just survived the main culmination of the third wave getting us into the information age [11]; and as I attempted to imply so far, it is reasonable to expect that a next future shock is coming. The question is, whether human race will or will not be washed away by the fourth wave of that high-tech intelligence, and eventually, who will be those potential survivors exactly.

In case of IA which is promoted in this paper, at least this could be held – *humans themselves would become their own successors* and *whatever injustice occurs would be tempered by our knowledge of our roots* [9, p. 19].

5 The Fourth Wave: Cyborgs

Obviously, there would be no point in discussing what it is like to experience the moment of our exhausted planet turning into a black hole singularity. We would be extinct even before that moment has come. Hence in this part, we continue focusing on the intelligence amplification conception and its roots and theoretical background.

As one can assume, the thought of improving human capability to gain experience and knowledge about the world more effectively is nothing new today. The ambition to increase not just our understanding of natural laws and formal rules but to enhance deliberately our own physical and intellectual power in general, moreover with the aid of technology, has been proposed strongly since the Enlightenment period of our history. In terms of human enhancement, Nicolas de Condorcet is often regarded as the most significant author for his posthumously published work *Esquisse d'un tableau historique des progres de l'esprit humain* (1795).[1] Especially in transhumanist literature (see for example [13–16]) we can find plenty of useful overviews with more detailed description of other retro-futuristic authors from the Enlightenment era to the twentieth century (e.g. Benjamin Franklin, Julien Offray de La Mettrie, Francis Bacon, Charles Darwin, Friedrich Nietzsche, Immanuel Kant, John Haldane, John Desmond Bernal, Julian Huxley etc.), therefore we will skip the historical digression and focus on more recent times.

Despite the fact that much inspiration for surpassing human physical and mental limits came from the field of philosophy and biology, it is no surprise that it was cybernetics where the idea of *intelligence amplification* (IA) appeared as an explicit scientific task, albeit not yet in its opposite role to AI. As a matter of fact, it came just after a series of famous Macy conferences that shaped in the 1940's and 1950's the field of cybernetics into a highly transdisciplinary domain. Biology, anthropology, physics, linguistics, mathematics, psychology, neuroscience, sociology and many others were involved deeply in defining goals, methods and language of the new discipline.

In his *Introduction to Cybernetics* (first published in 1956), Ross Ashby drawn his inspiration from a discipline of his own original expertise, i.e. psychiatry, and from medicine and biology as well, to scrutinize a way of how amplification of regulation and selection in general could work in the cybernetic manner. Quite by the same token as Butler, he was influenced by the theory of evolution. Based on this, he suggested to consider a *high power of appropriate selection* as *showing*

[1] In English published as *Outlines of an Historical View of the Progress of the Human Mind* in 1796 [12].

the behavioral equivalent of high intelligence. Such a brute conceptual reduction of intelligence to a mere selection skill was enabled by behavioral approach of course, but what is more important, it brought out the idea that *synthetically, consciously, deliberately* driven amplification of intelligence is possible and desired, be it human or artificial [17, p. 272].

Ashby was not alone with his thoughts. Expectations and hopes towards computer sciences were very high those days, and many others had no hesitation in proposing great progress in AI. Joseph Licklider proposed a *man-computer symbiosis* to come soon as our daily reality and persist for unpredictable time before computers will finally rule out the humankind, at least in all possible intellectual activities [18]. He envisioned the relation between man and machine as a kind of partnership in which human plays the first chair (so far) as an ultimate decision-maker aided in *real time* (which was really not possible in those days) by the information processing power of the computer. It cannot be denied that this became true and a matter of common fact for us at the beginning of the 21st century. Albeit, the characteristics (and constraints at the same time) of the computers, that, according to Licklider, made them complementary to humans, have been already overcome. Today's computers are not just capable of giving us answers in the real time, they are even able to predict what we are going to ask (at least Google works for me this way). They are not limited to process just one or very few elementary operations at a time anymore, and Licklider's concerns about *differences in speed and in language* are dissipating as we progress in automatic speech recognition, speech synthesis and user-friendly interfaces (e.g. ambient intelligence, intelligent environment, motion capturing, image recognition etc.). Indeed, we are already living this symbiotic lives together with our intelligent devices.

Andy Clark coined a term *systemic whole* [19, pp. 33–35] within his theory of *extended mind* to describe our mode of symbiosis with computers and various devices and artifacts we are using to touch and grasp the world. According to him, we actually externalize a lot of our mental and cognitive capabilities into those devices that are able to further amplify them, extend their meaning and significance, and at the same time we implement those devices and externalized processes back again into our identities in order to create the efficient unit that inhabits this universe. Clark calls us *natural-born cyborgs* [20] because according to him the externalization of our cognitive processes has started already with the use of language and with our ability to conceptualize the reality.[2] Thus we can say that the process of cyborgization actually started in the very same moment as the process of humanization of man.

Nevertheless, from a different point of view, it has not been always the case necessarily. Maybe we have not been always cyborgs as Donna Haraway proclaimed in her frequently cited *Cyborg Manifesto* [23]. Actually, the very first cyborg was a mouse before it was suggested by Nathan Kline and Manfred Clynes that human should cyborgize himself as well. The original purpose of

[2] Very close to this point of view was Doug Engelbart in his *Augmenting Human Intellect* from 1962 [21]. For detailed comparison of Clark and Engelbart see [22].

cyborgs was space exploration and it was meant again as taking an active role in the process of our own evolution [24, p. 345].

The mid-20th century was not just an era of artificial intelligence. It was the era of space race as well. Until 1969 it was very unsure whether man will ever step into fascinating but cruel and hostile universe with his own feet, and the poor mouse was thought to prove that it is feasible. In order to complete the space quest, man had to be re-designed on both physical and mental level. The key idea of Kline and Clynes was to create a homeostatic system assembled of properly modified human organism, technical devices and bio-chemistry control. Such system should be provided with sophisticated self-regulatory function (this is the moment when cybernetics comes to light again) that would guarantee constant homeostasis of the system even in the extraterrestrial conditions normally absolutely incompatible with human life. This space-adjusted man was named by Clynes *cyborg*[3] – a cybernetic organism, defined as *self-regulating man-machine system* or as *artificially extended homeostatic control system functioning unconsciously* [24, p. 347].

Very soon after *Cyborgs and Space* was published, Daniel Halacy provided us with probably the very first overview of cyborg (pre–)history, and moreover he elaborated sort of a *cyborg theory* as well. His *Cyborg – Evolution of the Superman* from 1965 [17] was foreworded by Manfred Clynes himself, and it goes far beyond the original proposal of cybernetic organism determined to adapt to the outer space. Besides an astonishing introduction into the history of various human body parts and organs replacements and reparations that served as injury or body disability compensation, Halacy considered (as the title of his book suggests) to use scientific knowledge to improve the current state of human body and mind – quite in the same manner as all the aforementioned philosophers who had accentuated the necessity of deliberate human enhancement. This way, Halacy linked together the idea of self-directed evolution of the human species and the idea of cyborgization.

Eventually, the first step onto the Moon was taken in 1969 by just a man equipped with technology, rather than by a real cyborg. Nevertheless, even though today we hardly ever think about cyborgs as space explorers, they have quickly become an archetypal personages of our own future and evolution. They are vehicles for our both techno-fetishist and techno-phobic tendencies. Of course, extreme positions always carry a potential of extremely bad decisions and effects, but under current conditions of our society being bound to its techno-culture, and due to my rather techno-optimist mind, I suggest to take Halacy's side:

> The old saying goes "if you can't beat 'em, join 'em". [26, p. 199]

After all, *cyborgs are extraordinary people – seeking extraordinary destinies* [26, p. 21]. Who could resist such an exciting fate?

[3] For the first time, the *cyborg* idea was published by the authors in *Astronautics* in 1960 [25], a year later a more deeply elaborated and more rigorous version of the paper was published in [24].

6 Conclusion

During the second half of the 20th century, cyborgs actually became a synonym for a superman. Today, we can basically identify cyborgs with bodily or cognitively enhanced humans who, on the theoretical level, are maintained by the concept of transhumanism; and on the practical level, there is a non-negligible support of human enhancement in more developed countries through their science policy emphasizing the so-called *converging technologies* paradigm. In brief, both transhumanism and technological convergence of particular scientific fields direct the scientists' effort towards making people more healthy, more happy, more social-skilled, high-perfomance and more intelligent. The discipline of artificial intelligence plays a crucial role in this process. An enormous amount of scientific progress is currently enabled by different kinds of artificial intelligence-based technologies and systems. They are also inevitable for cyborg technologies such as neuronal prosthesis, brain-computer interfaces, brain implants, extension and/or externalization of cognitive functions and so forth. It seems to be obvious that more or less specialized AI-based technologies are becoming important means of humans' self-transformation into a new species. After all, this is probably both more easier and definitely more desirable goal for the AI field, with great potential for humankind rather than for AGI-based superintelligence. This way, we are following actively the *Vinge's Survival Guide to Technological Singularity.*

By doing this, we challenge deep-rooted fundamentals of our daily reality represented by notions of naturalness and artificiality, corporeality, immortality, ego, individuality etc. At the first sight, it seems to be just a continuation of what we have started since we developed language and discovered magic of conceptualization.

Despite the seemingly problem-free situation, we suffer from lack of proper words to describe what our species is going through (or is wanted to) in relation to the current technohuman genesis. We tend to think about our today's ideas as being ground-breaking, but unfortunately it turned out in this paper that we apparently think about ourselves and our future in the very similar way for more than 150 years. What is desperately needed in order to keep up with the intelligence explosion is bouncing creativity that, according to Mikhail Epstein [27], should inhabit especially the academic field of humanities. The humanities are responsible for mental and cultural shifts in societies, and are the only ones with potential to guide humankind through such a significant turn in thinking and reflecting upon self. In order to fulfill such a role, we need creative thinking and a new discourse of humanities. At the end of the day, they are those taking care about human.

Critics of the idea of transhumanism, cyborgization, intelligent explosion, AI, IA and so on, make a point when doubting about feasibility of these ambitious projects. If we do not change our most inner and pampered attitudes, they will be right, and just a daydream will be left to us, no matter whether about AI or IA. We will keep writing down the same hopes and ideas again and again, but will never complete the quest of creating superintelligence. If we want to be direct ancestors of the supraintelligent humanoid species, we have to face this challenge and transform not just our bodies but our way of thinking in the first place.

References

1. Russell, S., Norvig, P. (eds.): Artificial Intelligence: A Modern Approach. Prentice Hall (2010)
2. Good, I.J.: Speculations concerning the first ultraintelligent machine. Advances in Computers 6, 31–88 (1966)
3. Searle, J.: Minds, Brains and Science. Harvard University Press, Cambridge (2003)
4. Warwick, K.: March of the Machines. University of Illinois Press (1997)
5. Butler, S.: Darwin among the machines. In: A First Year in a Cantebury Settlement with Other Early Essays, A. C. Fifield edn., pp. 179–185 (1914) (first published in 1863)
6. Kurzweil, R.: The Singularity Is Near. Viking, New York (2005)
7. Darwin, C.: On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. John Murray (1859)
8. Hayles, K.: How We Think: Digital Media and Contemporary Technogenesis. University of Chicago Press (2012)
9. Vinge, V.: The coming technological singularity: How to survive in the post-human era. In: Landis, G.A. (ed.) Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace, pp. 11–22. NASA Publication CP-10129 (1993)
10. Havel, I.: On the way to intelligence singularity. In: Kelemen, J., Romportl, J., Zackova, E. (eds.) Beyond Aaritifical Intelligence: Contemplations, Expectations, Applications, vol. TIEI 4, pp. 3–26. Springer (2013)
11. Toffler, A.: The Third Wave. William Morrow and company, Inc., New York (1980)
12. de Condorcet, N.: Outlines of an Historical View of the Progress of the Human Mind, Philadephia (1976)
13. Bostrom, N.: History of transhumanist thought (2005), `http://www.nickbostrom.com/papers/history.pdf` (cited December 12, 2013)
14. More, M.: The philosophy of transhumanism. In: More, M., Vita-More, N. (eds.) The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future, pp. 3–17. Wiley-Blackwell, Singapore (2013)
15. Hansell, G.R., Grassie, W. (eds.): H+/-: Transhumanism and Its Critics. Metanexus, Philadephia (2011)
16. Tirosh-Samuelson, H.: Eine auseinandersetzung mit dem transhumanismus aus jdischer perspektive. In: Coenen, C. (ed.) Die Debatte ber "Human Enhancement": Historische, Philosophische und Ethische Aspekte der Technologischen Verbesserung des Menschen, pp. 307–328. Transcript Verlag, Bielefeld (2010)
17. Ashby, R.: Introduction to Cybernetics. Chapman & Hall Ltd., London (1957)
18. Licklider, J.: Man-computer symbiosis. IRE Transactions on Human Factors in Electronics HFE-1, 4–11 (1960)
19. Clark, A.: Supersizing the Mind: Embodiment, Action, and Cognitive Extension. Oxford University Press, New York (2011)
20. Clark, A.: Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence. Oxford University Press, New York (2007)

21. Engelbart, D.: Augmenting Human Intellect: A Conceptional Framework. Prepared for: Director of Information Sciences Air force Ofce of Scientic Research, Washington (1962)
22. Zackova, E., Romportl, J.: Extended mind: Is there anything at all to be externalised? In: Kelemen, J., Romportl, J., Zackova, E. (eds.) Beyond Aaritifical Intelligence: Contemplations, Expectations, Applications, vol. TIEI 4, pp. 213–221. Springer (2013)
23. Haraway, D.: Simians, Cyborgs, and Women: The Reinvention of Nature. Free Association Books, London (1998)
24. Kline, N., Clynes, M.: Drugs, space, and cybernetics: Evolution to cyborgs. In: Flaherty, B.E. (ed.) Psychophysiological Aspects of Space Flight, pp. 345–371. Columbia University Press (1961)
25. Kline, N., Clynes, M.: Cyborgs and space, pp. 26–27, 74–76. Astronautics (September 1960)
26. Halacy, D.S.: Cyborg – Evolution of the Superman. Harper & Row Publishers, New York (1965)
27. Epstein, M.: The transformative Humanities: A Manisfesto. Bloomsbury Academic (2012)

Cyborg Tales: The Reinvention of the Human in the Information Age*

Jelena Guga[1,2]

[1] Department of Interdisciplinary Activities, New Technologies Research Centre
University of West Bohemia, Pilsen, Czech Republic
[2] 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 chapter 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. Furthermore, selected examples of the existing real-life cyborgs will show that SF narratives are not merely limited to the scope of imagination but are a constituent part of lived experience, thus blurring the boundaries between reality and fiction.

Keywords: cyborg, science fiction, cyberpunk, body augmentation, cognitive enhancement, artificial organisms, holograms, memory.

1 Cyborg Histories: Genealogical Overview of Origins and Meanings of Cyborg in Fiction, Science and Theory

Throughout the history, with every technological breakthrough, innovation or revolution, people have always imagined possible futures that new technologies at hand might bring about, for better or for worse. In our predictions and projections of hopes and fears onto the future, it is literature, art and film that have not only had an important role in shaping the ways we imagine the future of

* This work was supported by projects "Interdisciplinary Partnership for Artificial Intelligence" (CZ.1.07/2.4.00/17.0055) and "Excellence in Human Resources as a Source of Competitiveness" (CZ.1.07/2.3.00/30.0013), OP Education for Competitiveness, funded by the European Social Fund in the Czech Republic and by the State Budget of the Czech Republic.

J. Romportl et al. (eds.), *Beyond Artificial Intelligence,*
Topics in Intelligent Engineering and Informatics 9, DOI: 10.1007/978-3-319-09668-1_4

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, pp. 2–8], 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. It was a mouse with a Rose osmotic pump implanted under its skin, injecting chemicals into the 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 broader 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 chapter intertextually entertain a thought from the manifesto which states that "the boundary between science fiction and social reality is an optical illusion" [3, p. 149]. Through various examples coming from scientific research as well as from artistic practices, 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 grand-parents' have insufficient "now" to stand on. We have no future because our present is too volatile. [4, p. 40]

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 essay: 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 no longer 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 cyberpunk futuristic scenarios in order to better understand or figure out and cope with the technological cacophony of our present. So, for the purposes of this chapter, 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.

2 Digitally Mediated Cyborg Identities

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 and the ways we perceive what being human in an era of ubiquitous technologies means. 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 use of new technologies on global scale 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 day. The ubiquitous use of digital communication technologies has brought to the fore ambiguity, fluidity, contradictions and uncertainty when it comes to the ways we define our identities, our sense of self in relation to others, as well as our corporeal existence. Making an effort to cling to the "unified self" concept of Western philosophy in order to preserve a sense of certainty is futile. Such essentialist concept is unsustainable in technologically mediated multiple, networked realities we inhabit, where fixed or given identity turns out to be a mere illusion because in everyday life interactions we experience a distributed sense of self. In other words, our identities are subject to construction, multiplication, diffusion, fragmentation and change. The dynamics of interrelation between reality and virtuality is reflected on our corporeal existence and instead of Cartesian body-mind division, it is necessary to reexamine the role of embodiment in technologically mediated interactions and think of it in terms of inclusiveness and openness which can expand our faculties and capacities, rather than in terms of body denial we easily fall into.

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.), fog screens, and most recently, augmented reality (AR) devices such as Google Glass and other prototype eyewear alike, or wrist and armbands such as MYO. 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 fiction and among transhumanists as well) 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 which is already deeply caught up in the process of cyborgization? These questions are not posed to be given simple yes/no, good/bad essentialism-driven 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 which requires complex and risky surgeries, 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 through which one acquires new set of skills. Multiple gesturally driven image manifestations thus enable the expansion of bodily faculties and perceptual abilities.

In his novel *Idoru*, William Gibson entertained an 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 program 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 Hatsune's visual iconography. As a holographic celebrity, she performs in concerts with live musicians worldwide. These virtual constructs are not limited to the digital landscapes of cyberspace but exist in physical space as well. Moreover, 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: [they are] real in terms of having material effects on people's lives and playing a role in the formation of digital lifestyles, and [they are] fictional in insofar as [they] operate in conjunction with an elaborate fantasy narrative" [7, p. 106].

Apart from being a materialization of what Gibson has conceptualized in fiction, Idoru constructs can also be observed through the lens of Gilles Deleuze and Felix Guattari's philosophical concept of "body without organs" [8] in both metaphorical and literal sense. On the one hand Idoru are the hollow bodies, bodies of light 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 organ-ization. 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, single 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 Jean Baudrillard described it:

> 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 human-like 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. Rather, it is a question of what we are becoming through such intimate and intensive relations with our machines.

3 AI as the Paradigm for Cognitive Augmentation

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 the existing technologies under development. 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. Some of the most important issues in discussions on authenticity and simulation / original and copy, which are at the same time very often found in narratives of SF and cyberpunk films and literature, are the issues of consciousness, emotions and memory of artificially created organisms, the issues 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 research can be considered a foundation of an emerging science of forgetting. In the search for solutions to PTSD (Post-Traumatic Stress Disorder), anxiety disorder, addictive behaviors, 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, it is a malleable construct and 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 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. In other words, if protein synthesis is chemically inhibited prior to recollection of a memory, it 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 an emotional reaction, smell, words or looks. In other words, 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. However, 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 books. 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 advertising 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, the existential ground to stand on. 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 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 a 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 argues [15], expressing ourselves and at the same time reflecting upon ourselves in a strong, overwhelming and almost enchanting presence of such machines.

4 Present-Day Cyborgs: Body Enhancement in Scientific and Artistic Practices

Apart from the images of robotic and/or artificially grown beings discussed earlier, SF and cyberpunk 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 nail-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. However, these kinds of body modifications are not limited to the realm of science fiction only. It seem that Donna Haraway's claim about blurred boundaries between science fiction and social reality has proved to be true, considering that many of the concepts of science fiction are being materialized through scientific research and new media art practices, especially during the past two decades as they are becoming a part of our present experience. More and more, we see and hear of robotic, bionic and nano prostheses and implants, artificial tissues, genetically modified organisms, prenatal manipulation of gene structure, brain processes mapping and creating BCI (Brain-Computer Interfaces), and 3D printers that not only produce inert objects but can also print stem-cell based organ tissue or bone marrow. These are only a few of numerous examples which show that (bio)technological amplification of physical and mental faculties is no longer a mere product of science fiction – it has largely become a part of modern experience in which cyborgs represent a heterogeneous image of interplay between imagination and material reality constituted by science, art, and technology.

Technological bodily modifications are practiced today mostly for medical treatment purposes and prostheses and implants are used as a replacement of missing or dysfunctional body parts. 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 include scientific work of Professor Kevin Warwick who conducted experiments on his own body into which he implanted microchips described in the first chapter of this volume. Apart from purely scientific work, there is also a variety of artworks merging art and science, such as Stelarc's prosthetic bodily augmentations or bio and transgenic art projects by Eduardo Kac. In various scientific and non-scientific fields, researchers, artist, philosophers, techno-enthusiasts and homebrew cyborgs are all exploring the possibilities of expanding human capacities in order to overcome biological constraints and create more powerful, more resilient and more durable bodies which can adapt to the complex workings of the machines. In this sense, the restraining human body should be technologically enhanced in order to be able to adapt to technologically mediated environments where it would become an integral, compatible part of such eco-tech systems, as oftentimes depicted in science fiction.

Looking into many future scenarios found in science fiction narratives, Kevin Warwick poses a question of what has been accomplished so far in the fields of artificial intelligence, robotics and biomedicine, and what might the practical application of these technologies mean. In comparison to science fiction, he concludes that the rapid growth of scientific development has "not only done a catching up exercise but, in bringing about some of the ideas initially thrown up by science fiction, [it] has introduced wild card practicalities that the original story lines did not extend to" [18]. By conducting experiments which he has been an integral part of, subjecting his body to complex neurosurgeries, Warwick has brought the future of science fiction to the present of scientific reality. The first in a series of experiments involving digital identities, growing brains, deep brain stimulation, and human enhancement was his pioneering project in which an RFID (Radio Frequency Identification Device) was implanted into his upper left arm and enabled him to interact with the University's building: from controlling the lights and opening doors, to being greeted as he entered the front door. At about the same time Warwick had his chip implanted, Eduardo Kac, best known for his bio and transgenic art projects, did the same self-experiment called *Time Capsule*, only in a different context: instead of carrying out the procedure behind the closed door of a laboratory, Kac performed it in a gallery space open for the public. After he inserted the microchip into his left ankle, he used a scanner to activate its unique numerical code, which he then used to register himself in the US database for lost animals both as animal and owner under his name [19]. While the experiment conducted by Eduardo Kac can be read as an attempt to blur the distinction between species and between biological and artificial, Kevin Warwick's focus is more directed to introducing the technical possibilities. However, both experiments address the issues of identity, memory, safety, privacy, and surveillance in the digital culture.

Further research into the implications of augmenting the capabilities of humans led Professor Warwick to yet another more complex merger of biological and technological, i.e. human computer interface. With the aim to bridge the discrepancy imposed by the existing interfaces between technology and human motor and sensory systems, he has suggested direct interfacing with nervous system that could empower communication, expand multi-dimensional mental processing and memory as well as enhance bodily capabilities. In 2003, after being subjected to a complex two-hour neurosurgery, his nervous system was connected to a computer: "A stimulation current directly into the nervous system allowed information to be sent to the user, while control signals were decoded from neural activity in the region of the electrodes" [18, p. 14]. Having learned to distinguish the electro stimulation signals from those of his own nervous system, he carried out a number of bi-directional communication trials. Ultrasonic input enabled him to sense the distance of an object based on pulse frequency of nervous system. He was also able to teleoperate a robotic arm located at Reading University from a remote location (Columbia University in New York) and through feedback loop he could receive the signals from the fingertips of the robotic hand and sense the force it applied when manipulating objects. Other two trials included his wife who also got a chip implanted so their nervous systems could have a direct, sort of telegraphic communication. More precisely, his hand movement would send out a signal which her brain would receive in the form of pulses and vice versa. With such a direct brain to brain communication (even as rudimentary as it may be for now), Kevin Warwick has reconceptualized the notion of intimacy and (tele)presence. All the above experiments have proven to be practical and applicable for therapeutic medical purposes, but in the context of enhancing healthy individuals one cannot but wonder how far bodily and cognitive enhancement should be taken.

For more than three decades now, Australian artist Stelarc has been focusing on the dynamic relationships between body, technology and the future, while his works have been widely debated across theoretical and philosophical readings of digital environments, simulation, as well as human augmentation. Stelarc uses his own body as a medium in his artworks and performances and erases the distinction between the artist and the work of art through various cybernetic synergisms. Instead of being "a site for the psyche or the social," Stelarc sees the body as "a structure to be monitored and modified; the body not as a subject but as an object – not as an object of desire but as an object for designing" [20, p. 562]. In the processes of objectifying the body and technologically altering its functions through often very painful interventions, he has (literally) embodied robotics, micro-electronics, plastic surgery, biotechnologies, genetic engineering, AI, digital design, and communication systems, thus turning himself into a cyborg in various ways. His works and performances are based on the premise that the body is obsolete and that as such, it cannot keep up with the ongoing exponential technological development. Convinced that the body has to evolve together with technology, he made a series of art projects based on technological modification and augmentation of the body. Similar to what William Gibson does in fiction, Stelarc amplifies the present reality by pushing the limits of the body structure which must necessarily be reexamined, deconstructed and redesigned so that it could adapt to and expand within the technologically mediated environment.

His works such as *Suspensions*, *The Third Hand*, *The Stomach Sculpture*, *Fractal Flesh*, *Prosthetic Head*, *Exoskeleton*, and *The Extra Ear / Ear on Arm* [21] include collaboration with surgeons, engineers, scientists, etc., and demonstrate the unpredictable outcomes of biological and technological interweaving which requires new paradigms of body and identity. Throughout his projects, the artist's body underwent many invasive and non-invasive interventions: from piercing and stretching the skin, through expanding the body with external and internal robotic prostheses, to becoming a digitally created avatar governed by an AI. The body was also a part of a cybernetic feedback loop where it was controlled by the Internet data flow as well as by remote agents via the Internet, who teleoperated and initiated involuntary body movements of the artist, thus demonstrating that the body is not exclusively the site of the self – it can also function as a hollow body inhabitable by different agents and/or groups of agents.

As shown on the *Terminator* film franchise example discussed above, Stelarc has also evolved as a cyborg in accordance with the technological development and has gradually moved from hardware, to software, to wetware. For instance, after using the third robotic hand for several years, he started controlling it intuitively and perceiving it as a part of himself. However, this robotic prosthesis was an external one, meaning that it was put on and used on demand, mostly for the purposes of performances or similar public events. In other words, it was not permanently incorporated into the body's structure to be considered a constitutive part of artist's cyborg body, such as surgically implanted "eyeborg" for example, a prosthetic eyepiece which enables a colourblind artist Neil Harbisson to hear colours [22]. But unlike the mechanical third hand, Stelarc's *Extra Ear* was initially conceptualized as a permanent artificial ear on head, but due to possible risks and complications of such surgery, it has gradually developed into *Ear on Arm* project which represents a permanent change of body's architecture. The ear prosthesis was made of soft porous biocompatible tissue and was surgically attached to the artist's arm where it grew in and got intertwined with blood vessels and tissue around it. The ear was first multiplied through genetic engineering and then not only was it relocated, but its function was altered as well – instead of hearing, the ear emitted sound. This is an ongoing project that requires several more surgeries to get a 3D structure of the ear and make it an internet organ which can be accessed via the Internet so that everything in its immediate surrounding can be heard. Ear on arm is actually an experimental demonstration of technological replication, relocation and permanent function alteration of the body and/or body parts, and it transforms biological evolutionary architecture on the corporeal level. The volume and speed of information in technological and media environment we inhabit, supersede the notion of biological as a given and impose the need for "post-evolutionary strategies", to use Stelarc's term, which are being formed beyond philosophy and human physiology as we know it [20].

Of course, there are many other artists and art groups collaborating with scientists and engineers in order to create art projects that push the boundaries of body as a given and explore technological body modifications, while at the same time addressing and underlining problems as well as opportunities that lie in the existing information technology, robotics, nanotechnology, biotechnology, and medical technology. Amplification of physical and cognitive functions through symbiosis of technological and biological in artistic practices is marked by the term "bio-art" which was coined by Eduardo Kac while he was working on his *Time Capsule* project. The term bioart refers to those artistic practices which blur the distinction between art, science and fiction and in which prosthetics, implants, plastic surgery, genetic modification, etc. are used as the means or a medium of works of art. For example, French artist Orlan underwent a series of plastic surgeries to alter the architecture of the body, which she refers to as "carnal art" and defines it as "self-portraiture in the classical sense, but realized through the possibility of technology" [23]. In order to challenge the ideals of beauty, she has radically transformed her looks by combining different beauty standards imposed by famous paintings throughout history and applying them to her face – from turning herself into a hybrid of beauty, i.e. a mixture of Venus, Mona Lisa, Psyche, Diana and Europa in her project *The Reincarnation of Saint-Orlan*, to her *Self-Hybridization* series where she combined American-Indian, African and other non-western masks and sculptures with her own face that very often resulted in uncanny facial de-formations.

Other bio-artists such as Eduardo Kac, Critical Art Ensemble or Art Orienté Objet, to mention but a few, use not only the surface of the body as a medium of artistic expression but go further in exploring the options and possibilities that lie in manipulation of body fluids, tissue and DNA. These artists work in the field of biogenetic art which emerges from tight collaborations of artists and scientists. In both their theoretical and artistic works, art collective Critical Art Ensemble are specific for political activism when referring to new technologies and being subjected to their regulation. This is explicitly shown in their performance called *Flesh Machine* as well as in several other projects following it, where they researched new reproductive technologies or more specifically, donor programmes, and exposed the residue of eugenics in the fertility market thus addressing some of the crucial issues raised by biopolitics and bioethics debates. In the *2.6 g 329 m/s* project (named by the performance standard for bulletproof vests), Dutch artist Jalila Essaïdi explored the issues of safety and vulnerability in modern societies with high violence rate. Instead of creating bulletproof vests for protection, she created a bulletproof skin in collaboration with Forensic Genomics Consortium Netherlands and other relevant institutions. Bulletproof skin was genetically engineered by replacing keratin with spider silk protein, i.e. spider silk matrix produced by transgenic goat was implemented in human skin. Even though it didn't stop the full-speed bullet but only some of the partially slowed ones, the experiment was considered a success, especially for opening public debates on safety and, more importantly, on the topic of human enhancement and possible socio-political and cultural impacts the creation of bulletproof humans may have. The production of such skin can also be seen as a pioneering step towards creating what Natasha Vita-More envisions as "smart skin". Though she doesn't explicitly talk about bulletproof skin, it can be thought of as such given that she envisions "the syncretization of nanotechnology, biotechnology, information technology and cognitive and neuro science" for the purposes of engineering skin which could "repair, remake and replace itself" [24]. Despite the transhumanists' idealism when it comes to technological enhancement and especially to mind uploading coupled up with the belief in technologically aided human salvation of which I share none, I do agree with Vita-More's argument stating that we have long been integrating with other organisms such as bacteria for example, and that it is only natural to consider integration with other non-biological but biocompatible elements such as nanorobots or chemically charged agents which could result in mutually beneficial symbiosis. This statement echoes Haraway's theoretical stance on the boundary breakdowns between human and animal, animal-human (organism) and machine, and between physical and non-physical [3]. Additionally, examples can be found across a vast array of social practices including the production of artworks such as *Natural History of the Enigma* by Eduardo Kac, in which he created a new form of life, a genetically engineered hybrid consisting of the artist's DNA and flower petunia which he calls Edunia and which expresses his DNA in the red veins on pink flower petals. Blurring the boundaries between species is also found in the works of French artistic duo Art Orienté Objet (Marion Laval-Jeantet and Benoît Mangin), such as their 2011 project *May the Horse Live in Me* [25]. For the purposes of this performance, Marion's body was prepared for several months to build up tolerance to foreign bodies and was then injected with horse immunoglobulin which synthesized with her body, strongly affecting her body functions and nervous system. The basic idea behind this trans-species project was to try and determine what it is like to be a horse on experiential corporeal yet non-anthropocentric level.

The list of similar examples goes on as more and more artists engage in collaboration with scientists open to push the boundaries of established protocols in biotechnological research. Selected science-art projects presented above all question the relation between (bio)ethics and aesthetics and reestablish the role of art in everyday life. As noted by art and aesthetics theorist Miško Šuvaković, these artistic practices go beyond the scope of media or metamedia art and are defined as "postmedia art" given that they have a "specific function to mediate between cultural and social formations in historical and geographical actuality. ... Therefore, the ontology of these artworks is not aesthetical but social" [26, p. 114]. In other words, biogenetic, robotic and other forms of “postmedia art” which imply research and very often bold self-experiments in human augmentation conducted in collaboration with scientists, engineers, tattoo and piercing artists, ecologists, medical doctors, philosophers, psychologists, etc., all contribute to establishing a hybrid heterogeneous field which examines and questions the existing, present technological realities and reveal a myriad of future application potentialities. On the social level, they play a significant role in bringing what is going on behind the closed doors of laboratories to wider audiences, thus bridging the gaps between science, art and everyday life and more importantly, opening up public discussions on the uses of technology and its long term effects.

5 Conclusion

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 manipulation, 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, technological 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 our ever more transformative corporeal existence as well. 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. 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. Would those who chose not to upgrade or cannot afford technological enhancement be marginalized and considered obsolete and less competitive? What would be the optimal ways to regulate education, job market, healthcare system and other areas of life our very existence depend on? How will we 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? Being a cyborg should not be thought of as fictitious future of humanity for it is already interwoven into the fabric of our present. Therefore, it is necessary to step out of the radical oppositions of pros and cons of what future technologies may bring, but instead, create more nuanced, integrative and interdisciplinary theoretical and practical territories for further cyborgization processes which through various symbioses may enable us not only to live longer but prosper as well.

References

1. Gray, C.H., Mentor, S., Figueroa-Sarriera, H.: Cyborgology: Constructing the knowledge of cybernetic organisms. In: Gray, C.H. (ed.) The Cyborg Handbook. Routledge, London (1995)
2. Clynes, M.E., Kline, N.S.: Cyborgs and space. In: Gray, C.H. (ed.) The Cyborg Handbook. Routledge, London (1995)
3. Haraway, D.: A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century. 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 nad 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. Citadel, New York (1997)
12. Lehrer, J.: The forgetting pill. Wired (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. HarperCollins Publishers, London (1995)
18. Warwick, K.: Future issues with robots and cyborgs. Studies in Ethics, Law, and Technology 4 (2010)
19. Kac, E.: Time capsule (1997), `http://www.ekac.org/timcap.html`
20. Stelarc: From psycho-body to cyber-systems: Images as post-human entities. In: Bell, D., Kennedy, B.M. (eds.) The Cybercultures Reader. Routledge, London (2000)
21. Grzinic, M.: Stelarc: Political Prostheses & Knowledge of the Body. Maska MKC, Ljubljana (2002)
22. Bosker, B.: Cyborg Neil Harbisson on life with extra senses. Huffington Post (February 2013)
23. Orlan: Carnal art manifesto,
`http://orlan.eu/adriensina/manifeste/carnal.html`
24. Vita-More, N.: Nano-bio-info-cogno skin (March 2012),
`http://ieet.org/index.php/IEET/more/vita-more20120318`
25. Debatty, R.: Que le cheval vive en moi (May the horse live in me) (August 2011),
`http://we-make-money-not-art.com/archives/2011/08/que-le-cheval-vive-en-moi-may.php#.UcHH0-dpO8J`
26. Šuvaković, M.: Epistemologija umetnosti ili O tome kako učiti učenje o umetnosti (Epistemology of Art or How to Study Art Studies). Orion Art, Belgrade (2008)

Heteronomous Humans and Autonomous Agents: Toward Artificial Relational Intelligence

Hamid R. Ekbia

School of Informatics and Computing
Indiana University Bloomington, USA
hekbia@indiana.edu

Abstract. The notion of "autonomy" is a central concept and a generative metaphor in many AI approaches and systems. It also embodies a tension that is inherent to a persistent and sustained trend in AI that can be called "autonomist AI," whose objective is to build systems that are, on the one hand, complex and intelligent enough to initiate actions on their own, and, on the other, simple enough to be understandable and controllable by human beings. Tracing the origins of autonomist AI in some of the basic tenets of modernity, I show how the above tension is manifested in theories of affect, morality, and knowledge. I argue that these tensions arise largely because of adherence to a substantivist view, and propose a reversal to what I call Artificial Relational Intelligence.

Keywords: artificial relational intelligence, substantivism, relationalism, Eliza Effect, autonomous system, modernism, morality.

1 Introduction

On February 16, 2011, law enforcement officers in California arrested Chris Butler, also known as P.I. Mom, along with Norman Wielsch, a Narcotic Enforcement Team Commander, on charges of embezzlement, burglary, and conspiracy, as well as drug-related crimes. Butler, the founder of a company that provided service to women suspicious of cheating husbands, had signed a contract with *Lifetime Television* for a reality TV show on the same topic, where his alleged operations on behalf of betrayed wives were featured as true stories. His TV shows, as such, crossed the boundary between fiction and reality in a delicate manner, involving some of his real-world clients as actors. In parallel with this, in collusion with Wielsch, he was involved in illegal operations involving the resale of narcotics confiscated by law enforcement. In an interesting turn of events, which led to his arrest, TV and FBI cameras were capturing the same scene that was player out at once as real and as fantasy [1].[1]

[1] Butler's story is also narrated in detail on National Public Radio: http://www.thisamericanlife.org/radio-archives/episode/447/transcript, and it has its own dedicated Wikipedia page: http://en.wikipedia.org/wiki/Chris_Butler_private_investigator

J. Romportl et al. (eds.), *Beyond Artificial Intelligence,*
Topics in Intelligent Engineering and Informatics 9, DOI: 10.1007/978-3-319-09668-1_5

In an interview with reporters after his arrest, Butler mused, “The problem with people is that they want to believe you. You give them a little, and they take it from there.” What he calls a “problem”, however, is in fact a well-understood human trait with its own advantages and downsides. On the advantageous side, linguists, psychologists, and cultural anthropologists have shown that the capability of human beings to attribute beliefs and intentions to others is at the root of human sociality (see for example [2]). On the flip side, however, this trait can lead to attributions that go beyond the realm of the real, the sensible, or even the probable. This latter aspect is well known to AI practitioners who, as far back as 1976, came to notice the ubiquity of the *Eliza Effect* in human-machine interactions [3]. Initially considered as “the susceptibility of people to read far more understanding than is warranted into strings of symbols – especially words – strung together by computers” [4], the phenomenon seems to be much more general and prevalent, extending beyond strings of symbols to include actions, appearances, and affects. Ekbia in [5] proposes *Generalized Eliza Effect* (GEE) to refer to this broader phenomenon, and to show the penchant among AI practitioners to inadvertently use GEE to deal with tensions that often arise between their scientific and engineering aspirations, and to sometimes lure people into unrealistic claims about AI systems, committing in the process what he calls the *Attribution Fallacy.* However, as mentioned above, attribution has a light and positive side that also needs to be considered in AI, and this is what I would like to pursue here. The concept of “autonomy” provides a useful angle for this purpose.

This is how the article proceeds from here. We start with a brief examination of “autonomous systems” in AI and the way they are designed and represented in AI literature – in particular, the tendency in AI to erase the “supplements” that surround and enable the functioning of AI systems. Then, we trace the origins of this tendency in a substantivist philosophy, discuss its articulation in AI theories of affect, morality, and knowledge, and show the paradoxes that it faces in theory and practice. Finally, situating these paradoxes in the broader socio-historical developments of modernity, we explore the possibility of a reversal that could allow us to deal with them in a productive way.

2 Autonomous Systems in AI

The idea of “autonomy” as a generative metaphor has shaped a great deal of thinking and research in AI. A persistent trend, which can be called “Autonomist AI”, has survived paradigmatic changes in approach and technique. This trend is perhaps most explicitly represented in the research that is conducted on “autonomous agents”. The standard definitions of these agents describe them as follows:

> Autonomous agents are computational systems that inhabit some complex dynamic environment, sense and act autonomously in this environment, and by doing so realize a set of goals or tasks for which they are designed. [6, p. 108]

> An *autonomous agent* is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future. [7, p. 25]

Although somewhat different in details – e.g., in how they talk about the agent's "agenda" as something that is "its own" or for which it is "designed" – the common premise in both of these (and other definitions) is that they conceptualize autonomous agents as *inhabiting an environment*. However, as Elisabeth Wilson in [8] points out, "AI researchers have perhaps been more attentive, philosophically, to the autonomy that emerges for this agent ('its own agenda') than in how this autonomy has been and continues to be constituted through relations to a milieu ('within and part of an environment')."

In addition, to the extent that "environment" is featured in accounts of autonomous agents, it acquires a very narrow and often underdetermined character. That is, environment is understood as a set of features and properties that the agent *senses* and *acts upon*. What is often lost in this conceptualization is the fact that the environment also *supports* the agent in carrying out its actions [9]. This is a basic but important insight the flouting of which has serious implications for how we think about the world, about the nature of intelligence (natural or artificial), and about how we design systems [10].

Jacque Derrida makes this point using the playful rubric of "dangerous supplements" which, according to him, have two different significations: "A surplus, a plentitude, enriching another plentitude, the *fullest measure* of presence," or alternatively, "an adjunct, a subaltern instance which *takes-(the)-place* [*tient-lieu*]" [11, pp. 144–145]. In other words, that which supplements (e.g., environmental scaffoldings) can be simply a surplus that can be ignored but also as something that substitutes. Here Derrida is attracting our attention to the important role of supplements, the things that sometimes "take the place" of something.

Common practice in AI tends to focus on the first signification of supplements at the expense of the second one. This is best manifest in public portrayals of AI systems. A cursory comparison of pictures in Figure 1 shows, for instance, how the portrayal of the humanoid robot in the media as a solitary and "autonomous agent" erases the complex surrounding support structure provided by human beings, devices, and infrastructures, in the same fashion that media portrayals of chess contests between humans and machines take out of the picture the critical role of humans as surrogate players, spectators, analysts, and so forth, cf. [12]. The point, however, is more than the simple issue of media representation, and is symptomatic of a deep-rooted philosophy that can is referred to as "substantivism".

According to this view, an attribute or property P (e.g., intelligence, affect, morality, expertise, etc.) of the members of a social group G (humans, animals, robots, etc.) is the real and substantive possession of the members of that group [13]. Accordingly, the behaviors of the members of G can be largely determined by their membership in the group. An opposing view, which is sometimes

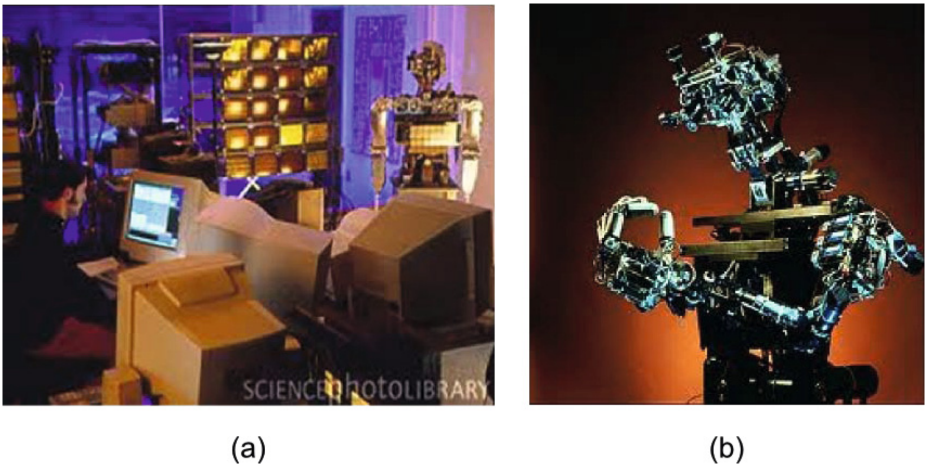

Fig. 1. The robot Cog with (a) and without (b) supplements

called “relational”, would posit that P is a set of attributions by the social group G – P *is in the eye of the beholder*, in other words. You are an expert on a topic such as medicine, for instance, if medical experts judge you as one. By the same token, you are “intelligent” if the members of the group of intelligent entities (humans?) judge you to be intelligent. This radical version of the relational view might be difficult to defend because, in its attempt to avoid the essentialism of the substantivist view, it puts an undue emphasis on outside relationships. A different version of the relational view, however, is conceivable, which would go like this: P is crucially dependent on the performative capabilities of other entities in G that are “outside” the individual. From this perspective, P *is in the act of performance and participation*; it is in the capability to interact and relate meaningfully with relevant others; or, to put it most succinctly, P *is a mediation* [14].

Both of these views – the substantivist and the relational (in its various versions) – can be traced in the history of AI, starting with its early origins in Alan Turing who speculated that “instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child’s?” [15]. Basically, Turing, advocating a relational view, wonders if by putting childlike AI systems in the right situation we might be able to turn them into recognizably intelligent beings among adult humans. In fact, in this reading, the Turing Test itself is a clear demonstration of the relational view, as it is organized around the notion of “performance”.

More recent incarnations of the relational view can be found in the interactionist thread in AI, which despite the best efforts of its advocates has not been able win a noticeable space in AI research and practice, giving way to the dominant substantivist view [16]. Many different parameters account for this course of development in AI: the philosophical tradition and the dominance of

Cartesian dualism, the development of computing and the prevalence of the von Neumann architecture, the individualist and sensationalist cultural milieu of the time, the largely militaristic funding structure in the US, close academic affiliation with cognitive psychology as opposed to social sciences, and so forth [5]. We do not have space to examine all of these parameters here. However, to understand the origins of that particular strand of the substantivist view that we called autonomist AI we need to step back and examine the historical context of AI developments as a modernist project.

3 Human Heteronomy and the Paradox of Modernism

Autonomy, as a kind of freedom, was introduced into modern philosophy in contrast to "negative" freedom – the right to do as one pleases, unimpeded by others [17]. It had to do with the willingness of citizens to surrender some of their rights as individuals and "think of their social membership as essential, not merely accidental, to who they are" [19, pp. 479–480]. As such, at its origin, autonomy was not conceived individualistically. In particular, it was not related to the ideal of autonomous judgment in the sense of deciding for oneself. Furthermore, the kind of moral freedom brought about by autonomy was considered by Rousseau to be superior to other natural or civic freedoms: "for the impulsion of mere appetite is slavery, and obedience to the law one has prescribed for oneself is freedom" [17, I.8, iii]. Given the insatiable character of human desires – something that separates us from all other animals – he finds this positive moral freedom to be the only way for humans to become "truly the master of himself [sic]" by submitting to a law that they themselves have prescribed.

The notion of self-prescribed law brings us to the etymology of the term – *auto* ("self") and *nomous* ("law") – and its inherent relationship to democracy as a form of governance. However, it also brings out an inherent tension and paradox in Rousseau's formulation, which commentators, philosophers, and modern societies in general have dealt with ever since. This paradox has to do with another fundamental feature of the human condition – namely, our dependence on others:

> Plants are fashioned by cultivation, and men by education. If man were born big and strong, his size and strength would be useless to him until he had learned how to use them; they would create prejudice against him; and, left to himself, he would die miserably before knowing his needs. We complain of the state of infancy; we do not see that, if man had not begun by being a child, the human race would have perished.[2] [18]

The emphasis that Rousseau puts on education and, more broadly, on the ongoing neediness and dependence of human beings on other fellow humans reveals a central dilemma of his thinking, but also a central paradox of modern

[2] The close parallel between Rousseau's view of the human condition and Turing's idea of a childlike AI system is interesting.

times – namely, the paradox of autonomy versus neediness, of voluntarism versus coercion:

> Rousseau's system is somewhat paradoxical ... The standard to which will must conform – is itself non-voluntaristic; therefore, contradictory. The standard that gives will its object is the very negation of voluntarism. [20, p. 121]

This neediness and dependence of human beings on others, coupled with inevitable asymmetries and social inequalities among individuals, creates the ground for social domination and the undermining of autonomy – hence, the notion of "dangerous supplement" that Derrida has playfully written about (see previous section). The danger calls for a political solution, of which Rousseau's *Social Contract* is a foundational articulation. Despite the danger, however, Rousseau does not counsel self-sufficiency because the independent and unattached beings of this imaginary scenario would be not only devoid of affect and love but also language, reason, and virtue – even selfhood itself [19, p. 487]. This is at the root of the paradox of Rousseau's system that attracts our attention to the opposite term *heteronomous*: subject to different laws, according to the Oxford Dictionary.[3]

Rousseau's concept has influenced political and moral philosophy, and also political thought, in the last three centuries, particularly the political ideology of various versions of liberalism that have come to play a dominant role in contemporary Western societies. Michel Foucault discusses the paradox of liberalism as having to do with the tension between freedom and coercion in liberal ideology. Freedom is the purported historical gift of liberalism to humanity. However, liberalism does not guarantee, provide, or even respect freedom. What liberalism purports to do is to produce what we need to be free – the conditions, organizations, instruments, etc. that create the possibility for the production of socio-economic, legal, and political freedoms. And the way it does this is by "the establishment of limitations, controls, forms of coercion, and obligations relying on threat, etcetera" [21, p. 64]. Consequently, freedom in liberal societies has to be constantly produced and accomplished; it is not a given. This duality between freedom and the coercive instruments and techniques for producing freedom led to what Foucault called the "paradox of liberalism".

In summary, the paradox of modernism finds various shapes and forms in modern societies, which seek to establish a social order that, on the one hand, provides individual freedom and autonomy, and, on the other, subject those same individuals to a collective will that is variously conceived and implemented depending on the dominant ideology and political order. In other words, modernity creates individuals who are autonomous in certain ways and heteronomous in other ways; it takes back with one hand something that it gives to the individual with the other.

[3] Immanuel Kant, in his moral philosophy, has a different interpretation of "heteronomous": acting in accordance with one's desires rather than reason or moral duty.

4 Autonomy and the Paradox of AI

Given this state of affairs, perhaps we can consider the general interest in "autonomy", especially in the context of AI research, as a projection of the modern humanity's desire for independence and freedom, on the one hand, and their despair in understanding their neediness and dependence. Whether or not we take it as a projection, one can discern a tension in AI, where practitioners seek to build systems that are intelligent, powerful, and autonomous, on the one hand, but that can also be intelligible, flexible, and predictable (if not controllable). At one level, therefore, the paradox can be seen in the tension between "intelligence" and "intelligibility" [12]. At another level, it can also arguably be seen in how AI research has dealt with specific topics such as affects, ethics, knowledge, and so forth. In dealing with the paradox, AI research has largely tilted toward a monadic understanding of these topics, underplaying their dyadic and social aspects. In the following discussion, I would like to illustrate how a substantivist philosophy underlies this bias in AI research.

5 Affect: Monadic or Dyadic?

Affects and emotions have been of interest to AI from early on, starting with Turing and continuing since then. Herbert Simon, for instance, emphasized the relationship between affect and cognition as early as 1960's [22]. Generally, both substantive and relational views on affects can be identified in AI, and in computing in general. While the former considers affects as discrete states, internal to the individual, and transmitted in a loss-free manner from people to others or to computational systems, the latter understands them as dynamic, culturally mediated, and socially constructed and experienced [23]. The substantive view is perhaps best represented in Donald Norman's model [24], which closely mimics Card et al's model [25] of human mind (see Figure 3). The relational view, on the other hand, is best captured in the idea of an affective loop that posits that affect flows in dyadic interactions between the individual and another person or system.

The relational view of affect has a long and established history in psychoanalysis, as well as social and developmental psychology, where the "self" is believed to emerge in interactions with others, and not as built-in. Feelings, according to Rosaldo [26], "are not substances to be discovered in our blood but social practices organized by stories that we both enact and tell." In a similar fashion, psychotherapy also seeks to build autonomy through relatedness. In fact, the practice of psychotherapy itself can be understood as a process through which autonomy is accomplished through the inter-subjective regulation of affect. This represents a paradox of psychotherapy [8, p. 85]:

> Ideally, psychotherapy builds not sovereign subjects but individuals who can both recognize their own self-states and modulate those states in relation to others. The inter-subjective regulation of affect is one of the

> means by which such autonomy emerges. In this sense, psychotherapy is an instance of a more general dynamic: all modes of autonomy are acquired affectively and relationally.

Child psychologists have similarly argued that infant development is a dyadic process. The famous proclamation by Donald Winnicott [27] speaks to this point: "There is no such thing as a baby" – i.e., a baby cannot exist alone, but is always and essentially part of a relationship. Along the same lines, Fonagy et al. [29, p. 4] argue as follows:

> The baby's experience of himself as an organism with a mind or psychological self is not a genetic given. It is a structure that evolves from infancy through childhood, and its development critically depends upon interaction with more mature minds, who are benign and reflective in their turn.

These observations lead to a dyadic notion of affect that operates through what can be called an "affective loop", which is at work even in those cerebral situations that seem to be far removed from emotional attachment – e.g., in championship chess. For instance, Kasparov's defeat by the Deep Blue can be equally attributed to the power of the machines as to his failure to establish an affective loop with the opponent (Figure 2d). As Wilson [8] points out:

> Kasparov's customary tactics of intimidation aren't simply a projection onto the opponent – a kind of one-sided attack. Rather, Kasparov, when he is most effective, recruits his opponents into an affective intimacy, albeit intimacy rooted in fear. The pertinent issue is not the emotion *in* Kasparov (Is he angry? Is he afraid?), as if he operates as an affective monad (an isolated talent); rather it is the emotional relationality between Kasparov and his opponent that governs, in part, whose intelligence will prevail.

One can, indeed, read Turing in a similar manner. Rather than emphasizing particular affects such as fear, joy, or anger, and their instantiation in machines, he is interested in how affectivity cultivates relationships between agents (specifically between humans and machines):

> Without much in the way of available theory about the development of mind, Turing nonetheless seems to intuit that interiorities (human and artificial) are built mutually, intersubjectively ... At important junctures [e.g., the biographical anecdote reported by [28] in the conversation between Turing and his partner Arnold Murray about dreams], Turing imagines thinking and feeling to be *chiasmatically* related rather than opposed and disjunctive. [8, p. 21]

Against these proposals and precursors, the approach of Autonomist AI to the question of affects and emotions has been largely monadic, seeking to inscribe and attach affects to individual artifacts, which leads to a neglect of the dyadic character of affect and the intersubjective aspects of emotion. Generally, in AI research, emotions *reflect* drive states but do not have much motivational force by themselves. This is, for instance, how the affective mechanisms of the "baby" robot Kismet are described by its designers [30, p. 55]:

> When in the homeostatic regime, a drive spreads activation to those [emotional] processes characterized by positive valence and balanced arousal. This corresponds to a "contented" affective state. When in the under-stimulated regime, a drive spreads activation to those processes characterized by negative valence and low arousal. This corresponds to a "bored" affective state that can eventually build to "sorrow". When in the overwhelmed regime, a drive ...

In brief, Kismet operates in three regimes, where different levels of arousal give rise to appropriate emotional states. While quite novel, the approach in the design of Kismet is based on a theory of drives and affects, which is in contrast with the views of psychologists such as Sylvan Tomkins, who also happened to have had a long-term interest in social and affective robots, long before AI projects such as Kismet came into being. According to Tomkins, our behaviors are largely regulated by affects, which are sustained and general in character, as opposed to drives, which are spatially and temporally specific and hence weak in motivating behavior. Affects, as such, take priority over drives. The hunger drive, foundational to behaviorism and also to Freud's theory of sexuality, for instance, is not powerful by itself. It becomes urgent (and so able to compel behavior) when it is amplified by, say, distress or enjoyment. It can similarly be attenuated or blocked by disgust or fear. Sexual drive is similarly diminished by shame, fear, apathy, or surprise. Humans act "not only by a responsiveness to drive signals but by a responsiveness to whatever circumstances activate positive and negative affect" [31, p. 22]. Therefore, "the creation of a humanomaton would require an affect system," according to [31, p. 18], not a drive system as we saw in Kismet.

Kismet, however, suffers from a more serious shortcoming in terms of its affective behavior. Its affective states, expressed in nine different facial expressions (happy, sad, angry, etc.), lack an important intersubjective affect that operates in the space between aspiration and dependency – namely, "shame". A negative affect triggered when positive affects of interest and enjoyment are obstructed (I want, but ...), shame brings to halt facial communication with the eyes down, head down, and blushing. Many psychologists and cultural anthropologists have argued that shame is the quintessential intersubjective and dyadic affect, which regulates social behavior – an aspect that should have been of natural interest to the designers of Kismet.

The dominance of drives instead of affects and the absence of shame have led Wilson to conclude that:

> In much of the Kismet project, internal states are deduced rather than felt, intellectually discerned rather than sympathetically known. Mutuality is executed rather than sensed ... The Kismet project is drawn to emotion, but then loses its nerve. [8, p. 55]

> ... one thing we might be able to declare on examining an autonomous robot like Kismet is that there is no such thing as an autonomous robot. [8, p. 74]

6 Morality

Similar issues in the area of morality and machine ethics, which is concerned with ensuring that the behavior of machines toward human users, and perhaps other machines as well, is ethically acceptable. Examples of research where this might be an issue are autonomous robots (social and home-based robots), autonomous driving, and autonomous drones and armed vehicles. In dealing with these topics, we also see the substantive and relational views in opposition. A substantivist sees morality as the outcome of individual reasoning, embodied in implicit or explicit ethical agents that are programmed to behave ethically, respectively, without an explicit representation of ethical principles or based on the calculation of the best action in ethical dilemmas using ethical principles [32]. A relational view, on the other hand, considers morality as the mutual accomplishment of individuals embedded in worlds of social values. From this perspective, guilt and shame are seen as moral affects necessary to constrain the individuated self from dangerous and asocial acts of impulse, lust, and violence [26].

The substantivist view has led some researchers to contend that "machines with a level of autonomy requiring ethical deliberation are already with us, and that their number and level of autonomy are likely to increase." Therefore, "the liability already exists, and machine ethics is necessary as a means to mitigate it" [33]. This leads these authors to the rather utopian idea that the ultimate goal of machine ethics is to create a machine that is an explicit ethical agent – a "Humans 2.0", which would be a better version of human beings:

> Machines, though, might have an advantage over human beings in terms of behaving ethically ... human beings, as biological entities in competition with others, may have evolved into beings with a genetic predisposition toward unethical behavior as a survival mechanism. Now, though, we have the chance to create entities that lack this predisposition, entities that might even inspire us to behave more ethically.

Interestingly, the substantivist view behind the AI approach allows these authors to make the more general essentialist claim about human beings as genetically disposed toward unethical behavior – a claim that would not sit well with those who expect a more balanced and nuanced image of human morality. Ironically, the same substantivist view gives rise to dystopian fears of a "post-biological" future, where "The human race has been swept away ... usurped by its own artificial progeny" [34]. It is, indeed, hard not to read the resonances of an ancient mythology in these images. Commenting on the related case of HAL in Arthur Clark's celebrated fiction, Bloomfield & Vurdubakis [35] wonder:

> For what is HAL's crime but the Original Sin? Moderns, having created thinking machines in their own image, immediately expect that these machines will – just like they themselves did – attempt to usurp the powers of their creators ... It is perhaps paradoxical but not unexpected that AI, the enterprise that is said to epitomize the workings of reason, is at the same time so heavily mythologized.

With these underpinnings, it is indeed hard to find work in machine ethics that could be considered a project in relational ethics. By way of contrast, however, it might be useful to point out Tomkins's vision on morality [31, p. 216]:

> Just as contempt strengthens the boundaries and barriers between individuals and groups, and is the instrument par excellence for the preservation of hierarchical, caste and class relationships, so is shared shame a prime instruments for strengthening the sense of mutuality and community whether it be between parent and child, friend and friend, or citizen and citizen. When one is ashamed of the other, that other is not only forced into shame but he is also reminded that the other is sufficiently concerned positively as well as negatively to feel ashamed of and for the other.

7 Knowledge and Expertise: Tacit and Explicit

As a final illustration of the contrast between substantivist and relational views, one can mention AI models of knowledge-intensive or expert systems. A substantivist view starts with the premise that knowledge can be explicitly captured in propositional statements connected by logic-based rules, whereas from a relational perspective tacit knowledge is the linchpin of human cognition. The question of tacit knowledge is probably most widely discussed by social scientists such as sociologists of science. Influenced by Wittgenstein's philosophy of language and Michael Polanyi's work on "tacit knowledge", many of these commentators emphasize that a significant part of human knowledge is not directly accessible to conscious thought. More recent work categorizes tacit knowledge into different types, showing the intricacies of acquiring, sharing, and even talking about knowledge. Collins and Evans [13], for instance, introduce what they call the "Periodic Table of Expertise". Although the term itself is not perhaps suitably chosen (there is nothing "periodic" about types of expertise), there are interesting insights in the work, some of which have direct bearing for AI and for our discussion. One key insight is in the idea of "interactional expertise", which has to do with the capability of talking in the language of a specialism in the absence of expertise in its practice. This kind of expertise, according to Collins and Evans, is quite ubiquitous, and can be found in many different situations and places – e.g., journalists or ethnographers studying a specialty; managers evaluating a specialist; peer-reviewers commenting on the merits of a manuscript or grant application; and so on.[4] In successful scenarios of these situations, there is usually a progression from "interview" to "discussion" to "conversation" between the contributing expert and the interactional expert, throughout which the latter incrementally learns how to speak the language of the former.

This account of interactional expertise brings it one step closer to the possibility of AI and expert systems with some kind of access to tacit knowledge, compared to alternative views based on a more strict understanding of embodiment (e.g., [37]). However, it still leaves a huge gap for systems such as Cyc [36] that have been built on the premise of rule-based encoding of explicit knowledge. Cyc is one of the longest running AI projects, which started in the early 1980's with the idea of encoding all encyclopedic knowledge, then shifted to what is called commonsense knowledge, and more recently it seeks to use semantic web techniques to pool the web as a source of knowledge. While this last move makes sense, given what we know about the tenets and principles of Cyc [5], it is unlikely that the project can come even close to its alleged goal of becoming the indispensible knowledge platform of computing or the semantic web. One of these tenets has to do with what can be called an autonomist epistemology, which considers knowledge to be the possession of an individual agent or system.

By way of contrast, Watson developed by IBM works on a hypothesis-driven model, and given what we know about its architecture it seems to come closer to the mark in terms of relationality, although it is too early to make any judgments in terms of its direction [38]. In particular, the disembodied character of the system and its reliance on explicit representation poses serious challenges – e.g., in breaking the meaning barrier [39].

8 Looking Ahead

To summarize what we have discussed so far, Autonomist AI – the prevailing perspective based on a substantivist view of intelligence, knowledge, affect, morality, etc. – faces serious challenges in its goal of creating autonomous systems. In particular, it deals with an inherent paradox that manifests itself differently in different systems: developing affective agents regulated by drives, moral agents without shame, expert systems with no tacit knowledge, and so forth. We traced the historical origins of this paradox in the modernist project that seeks to create autonomous human beings endowed with freedom but constrained by their inherent neediness and their social obligations. AI research doubly augments this paradox through its techno-cultural imaginary. In its attempt to carry out

[4] I need to add that Collins and Evans present their account of expertise as a "substantive" theory, and in opposition to relational views. In my view, this is why their account runs into problems and dilemmas that they acknowledge [13, p. 76]. More relevant to our purposes, however, what they consider "relational" is pure attribution, which is not the stance that I advocate here.

"[Modernity's] demiurgic ambition to exorcise the natural substance of a thing to substitute a synthetic one"[40], AI puts the modernist paradox in high relief.

Paradoxes, of course, do not have solutions. A productive strategy for dealing with paradox, as such, is "reversal" – that is, a strategy that flips the center and the margin, and thereby expands our understanding [41]. In the case at hand, Autonomist AI, based on a substantivist view, puts individual properties and agendas of agents at the center, pushing "supplements" and dyadic relationships with others and the environment to the margins. A strategy of reversal would, therefore, put these at the center, and de-emphasize individual attributes.

I propose such a reversal with the aim of pursuing what can be called Artificial Relational Intelligence (ARI). ARI is a way of thinking about AI so as to make it more realistic and more humble, but no less interesting and challenging in its aims than current alternatives. This is a topic that I hope to pursue in later writing. To reiterate the power of participation in human behavior, though, I would like to end with a comment on the opening story of this article [1]:

> Everyone in that room broke the law, but they were celebrating it like a civic duty. That's how Chris allowed them to see it. Because if Chris had a genius for fantasy, it was that he understood that everyone had their own particular fantasy, and he could spot it and harness it, and weave it together with the rest of the people in his web ... The Moms wanted to be on TV. Norm wanted to feel powerful again. The media wanted a good story. The Candyman got a little fantasy date. Even Carl told me that before he first blew the whistle on Chris, he hesitated. Not just because he was scared, but because he, too was taken by Chris's grand vision.

References

1. Berton, J.: Dirty DUIs' figure gets 8-year sentence (September 2012), http://www.sfgate.com/crime/article/Dirty-DUIs-figure-gets-8-year-sentence-3893977.php#ixzz29Z4eYLvc
2. Levinson, S.C.: Cognition at the heart of human interaction. Discourse Studies 8(1), 85–93 (2006)
3. Weizenbaum, J.: Computer Power and Human Reason: From Judgement to Calculation. W. H. Freeman & Co. (1976)
4. Hofstadter, D.R.: Fluid Analogies Research Group: Fluid Concepts And Creative Analogies: Computer Models Of The Fundamental Mechanisms Of Thought. Basic Books (1995)
5. Ekbia, H.R.: Artificial Dreams: The Quest for Non-Biological Intelligence. Cambridge University Press (2008)
6. Maes, P.: Artificial life meets entertainment: Life like autonomous agents. Communications of the ACM 38(11), 108–114 (1995)
7. Franklin, S., Graesser, A.: Is it an agent, or just a program?: A taxonomy for autonomous agents. In: Müller, J.P., Wooldridge, M.J., Jennings, N.R. (eds.) ECAI-WS 1996 and ATAL 1996. LNCS, vol. 1193, pp. 21–35. Springer, Heidelberg (1997)
8. Wilson, E.A.: Affect and Artificial Intelligence. University of Washington Press (2010)

9. Vygotsky, L.S.: Mind in Society: The Development of Higher Psychological Processes. Harvard University Press (1978)
10. Agre, P.E.: Computation and Human Experience. Cambridge University Press (1997)
11. Derrida, J., Spivak, G.C. (trans.): Of Grammatology. The Johns Hopkins University Press (1976)
12. Suchman, L.: Human-Machine Reconfigurations: Plans and Situated Actions. Cambridge University Press (2007)
13. Collins, H., Evans, R.: Rethinking Expertise. University Of Chicago Press (2007)
14. Ekbia, H.R.: Digital artifacts as quasi-objects: Qualification, mediation, and materiality. Journal of American Society for Information Science and Technology 60(12), 2554–2566 (2009)
15. Turing, A.M.: Computing machinery and intelligence. Mind LIX(236), 433–460 (1950)
16. Agre, P.E.: Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI. In: Bowker, G., Star, S.L., Gasser, L., Turner, W. (eds.) Social Science, Technical Systems, and Cooperative Work: Beyond the Great Divide. Psychology Press (1997)
17. Rousseau, J.-J.: The Social Contract and other Later Political Writings. Cambridge University Press, Cambridge (1997)
18. Rousseau, J.-J.: Emile, or On Education. Basic Books, New York (1979)
19. Neuhouser, F.: Jean-Jacques Rousseau and the origins of autonomy. Inquiry 54(5), 478–493 (2011)
20. Riley, P.: Will and Political Legitimacy: A Critical Exposition of Social Contract Theory in Hobbes, Locke, Rousseau, Kant, and Hegel. Harvard University Press (1982)
21. Foucault, M.: The Birth of Biopolitics: PLectures at the College de, pp. 1978–1979. Palgrave McMillan, London (2008)
22. Simon, H.A.: Motivational and emotional controls of cognition. Psychological Review 74(1), 29–39 (1967)
23. Boehner, K., Depaula, R., Dourish, P., Sengers, P.: Affect: From information to interaction. In: Critical Computing Fourth Decennial Aarhus Conference, pp. 59–68 (2005)
24. Norman, D.: Emotional Design: Why We Love (or Hate) Everyday Things. Basic Books (2004)
25. Card, S.K., Moran, T.P., Newell, A.: The Psychology of Human-Computer Interaction. Lawrence Erlbaum Associates (1983)
26. Rosaldo, M.: Towards an anthropology of self and feeling. In: Shweder, R.A., Le Vine, R.A. (eds.) Culture Theory: Essays on Mind, Self, and Emotion. Cambridge University Press (1984)
27. Winnicott, D.W.: The Family and Individual Development. Basic Books (1966)
28. Hodges, A.: Alan Turing: The Enigma. Random House, London (1983)
29. Fonagy, P., Gergely, G., Jurist, E., Target, M.: Affect Regulation, Mentalization, and the Development of Self. Other Press (2002)
30. Breazeal, C., Scassellati, B.: Infant-like social interactions between a robot and a human caregiver. Adaptive Behavior 8(1), 49–74 (2000)
31. Tomkins, S.S.: Affect, Imagery, Consciousness. Springer Publishing Company (1963)
32. Moor, J.H.: Why we need better ethics for emerging technologies. Ethics and Information Technology 7(3), 111–119 (2005)

33. Anderson, M., Anderson, S.L.: Special issue on machine ethics. IEEE Intelligent Systems 21(4) (2006)
34. Moravec, H.: Mind Children: The Future of Robot and Human Intelligence. Harvard University Press, Cambridge (1988)
35. Bloomfield, B.P., Vurdubakis, T.: IBM's chess players: On AI and its supplements. The Information Society 24(2), 69–82 (2008)
36. Lenat, D.B., Witbrock, M.J., Baxter, D., Blackstone, E., Deaton, C., Schneider, D., Scott, J., Shepard, B.: Harnessing Cyc to answer clinical researchers' ad hoc queries. AI Magazine 31(3), 13–32 (2010)
37. Dreyfus, H.L.: Why Computers Must Have Bodies in order to Be Intelligent. Review of Metaphysics 21(1), 13–32 (1967)
38. Ferrucci, D.A., Brown, E.W., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J.M., Schlaefer, N., Welty, C.A.: Building Watson: An overview of the DeepQA project. AI Magazine 31(3), 59–79 (2010)
39. Mitchell, M.: Artificial intelligence and the barrier of meaning (2012), http://fora.tv/2012/10/14/Melanie_Mitchell_AI_and_the_Barrier_of_Meaning
40. Baudrillard, J.: Simulacra and Simulation. University of Michigan Press (1983)
41. Derrida, J., Bass, A. (trans.): Margins of Philosophy. University Of Chicago Press (1982)

Moral Enhancement and Artificial Intelligence: Moral AI?

Julian Savulescu and Hannah Maslen

Oxford Uehiro Centre for Practical Ethics
University of Oxford, UK
{julian.savulescu,hannah.maslen}@philosophy.ox.ac.uk

Abstract. This chapter explores the possibility of moral artificial intelligence – what it might look like and what it might achieve. Against the backdrop of the enduring limitations of human moral psychology and the pressing challenges inherent in a globalised world, we argue that an AI that could monitor, prompt and advise on moral behaviour could help human agents overcome some of their inherent limitations. Such an AI could monitor physical and environmental factors that affect moral decision-making, could identify and make agents aware of their biases, and could advise agents on the right course of action, based on the agent's moral values. A common objection to the concept of moral enhancement is that, since a single account of right action cannot be agreed upon, the project of moral enhancement is doomed to failure. We argue that insofar as this is a problem, it is a problem for some biomedical interventions, but an agent-tailored moral AI would not only preserve pluralism of moral values but would also enhance the agent's autonomy by helping him to overcome his natural psychological limitations. In this way moral AI has one advantage over other forms of biomedical moral enhancement.

Keywords: moral enhancement, artificial intelligence, moral psychology, cognitive biases, biofeedback.

1 Introduction

Human moral psychology is limited by design in many ways. It is subject to biases that lead us to care excessively about close acquaintances in comparison to strangers, to care more about what happens in the near than the distant future, to empathise with individuals but not groups, and to distrust and even wish harm on those whom we perceive as outside our racial, national or cultural groups. Far from proceeding in the rational and deliberative way we might hope, most of our moral views and decisions are based on immediate intuitions, emotional responses, and gut reactions. Reasoning, if it even comes in to the picture, is often used merely to rationalise what we intuitively believed anyway.

As the world becomes increasingly technologically advanced and increasingly globalised, the consequences of human moral limitations become more profound: whilst our moral psychology evolved to be effective in small groups, many modern

J. Romportl et al. (eds.), *Beyond Artificial Intelligence,*
Topics in Intelligent Engineering and Informatics 9, DOI: 10.1007/978-3-319-09668-1_6

problems – such as climate change and scarcity of resources – present global (not local) collective action problems. Human beings did not evolve to deal with such large-scale socio-moral challenges and pursuing some sort of moral enhancement will therefore be a crucial counter-measure.

So far, academic interest in human moral enhancement has tended to focus on biomedical interventions (such as genetic selection, pharmaceuticals and brain stimulation), often comparing them with traditional interventions (such as education and religion) (in particular, see [1–3]). There is scientific evidence emerging to suggest that there may be some biomedical interventions that have short-term effects consonant with commonly accepted morality (for example, by making people less xenophobic [4]). However, the effects of such interventions tend to be short-lived and, often, conceptualising a particular effect as a moral enhancement is contingent on accepting a particular – and often debatable – set of moral values. In this chapter we suggest a third potential mechanism for moral enhancement which we believe should be explored alongside traditional and biomedical interventions.

Following developments in pervasive computing and ambient intelligence, we propose that moral artificial intelligence (moral AI) could be developed to help agents overcome their natural psychological limitations. The moral AI would monitor physical and environmental factors that affect moral decision-making, would identify and make agents aware of their biases, and would advise agents on the right course of action, based on the agent's moral values. In being tailored to the agent, the moral AI would not only preserve pluralism of moral values but would also enhance the agent's autonomy by prompting reflection and by helping him overcome his natural psychological limitations.

2 Human Moral Limitation and Evolutionary Psychology

The moral psychology of the human species evolved to make them fit to live in small, close-knit communities. There were many reproductive and survival benefits to be gained from living in such communities; benefits such as communal childcare, pooling of resources like food, and better protection from predators. However, many of these benefits could only be obtained if individuals were disposed towards cooperation, presenting our ancient ancestors with recurring social dilemmas: individuals would be worse off if they act against their immediate self-interest (e.g. if they give away some of their resources) but everyone would be worse off if all the individuals act according to their self-interest. Thus, morality – at its most basic level which involves sharing and cooperating within a group – evolved to facilitate cooperation that promoted inclusive fitness (see [5]). Whilst this morality was clearly effective in the social context in which it evolved, it has various limitations which were either positively selected for, or simply were not detrimental for most of our evolutionary history.

For example, whilst we developed some disinclination to free-ride due to the longer term detrimental consequences of failures of cooperation, the efficacy of the brakes on free-riding are dependent on the group size being relatively small.

As groups become larger, individuals are more anonymous and free-riding becomes harder to detect. The moral disposition to cooperate for the benefit of the group diminishes as group size increases (for discussions of the free-rider problem in evolutionary theory see [6, 7]).

Due to evolving in small, tight communities with technology that could only affect our ancestors' immediate surroundings, the moral psychology that evolved is predominantly "myopic". It biases us to care much more about people in our immediate neighbourhood and in the immediate future than those people further away in space or time. Further – in line with the theory of kin selection – our altruistic propensities tend to demonstrate partiality towards family and friends. Because close relatives of an organism share some identical genes, a gene can increase its evolutionary success by promoting the reproduction and survival of related or otherwise similar individuals (see [8]). Further, a disposition to do favours that extends indiscriminately to strangers would expose us to too great a risk to be exploited by free-riders.

Correspondingly, our morality also evolved to engender a distrust of strangers. Where we do not know whether another individual will defect in a social dilemma situation, the costs of a false negative – believing that someone will cooperate when they in fact defect – are often higher than the costs of a false positive – believing that someone will defect when in fact they would have cooperated. The so-called negativity bias (whereby negative information about a potential exchange partner is more salient and produced more arousal than positive information) evolved as a defence mechanism where costs of false negatives are particularly high (see [9, 10]).

Further, the creation of clear group boundaries to signal whom one should direct prosocial behaviour towards would have been beneficial from the perspective of inter-group competition (see [11, 12]). However, the flip side of this strong group identification and attachment involves strong resentment of members of out-groups, resulting in antisocial behaviour and xenophobia.

It might be hoped that, although our morality evolved in what we now see as a limited way, we overcome these limitations when we deliberate about what to do in given situations – when we weigh pros and cons and decide on a course of action. However, recent science suggests that systematic moral reasoning is a rare phenomenon and often an illusion. Instead, most moral decisions are based on immediate intuitions, emotional responses, and gut reactions. Reasoning, if it even comes in to the picture, is often used merely to rationalise post-hoc what we intuitively believe anyway (see [13]).

3 The Grave Consequences of Moral Limitation

The consequences of our limited moral psychology are becoming increasingly profound. What was adaptive for most of our evolutionary history no longer equips us sufficiently to deal with modern global and technological challenges. In fact, some of the biases we developed are now mal-adaptive. The developments of science and technology have radically changed our living conditions and what we are capable of. We now live in societies with millions of citizens and with an advanced scientific technology which enables us to exercise an influence that extends all over the world and far into the future. This is leading to increasing environmental degradation and to harmful climate change. The advanced scientific technology has also equipped human beings with nuclear and biological weapons of mass destruction which might be used by states in wars over dwindling natural resources or by terrorists. Liberal democracies cannot overcome these problems by developing novel technology.

The potential to actively harm is not the only problem facing our globalised world. There is also a serious failure to aid those in need. In 2008 only 5 countries (Sweden, Luxembourg, Norway, Denmark, and the Netherlands) had reached the modest UN of aid amounting to 0.7% of GNP. The average for OECD-nations is 0.47% and the two biggest world economies, the US and Japan, lie at the bottom, at around 0.2% (see [14]). Further, individuals are naturally disinclined to act to avert environmental crisis. As the number of agents involved in a cooperation problem grows, the contribution of each agent to the total outcome becomes negligible or imperceptible. This leaves agents with little altruistic or even self-interested reason to contribute to solutions.

Elsewhere, one of us has argued that what is needed is an enhancement of the moral dispositions of citizens to extend their moral concern beyond a small circle of personal acquaintances, including those existing further in the future (see [15, 14]). The expansion of our powers of action as the result of technological progress must be balanced by a moral enhancement on our part. Otherwise, our civilisation is itself at risk. It is doubtful whether this moral enhancement could be accomplished by means of traditional moral education. There is therefore ample reason to explore the prospects of moral enhancement.

4 The Project of Moral Enhancement

Moral enhancement is the project of trying to improve moral cognition, motivation and behaviour. So far in debates about moral enhancement, attention has mostly been directed to biomedical means of affecting moral improvement. The results of scientific experiments are beginning to provide indication that some moral enhancement by biomedical means might be possible.

Empirical research has shown that the anti-depressant propranolol can reduce implicit racial bias [4] and produce less utilitarian judgement [16]. Work by Niels Birbaumer and colleagues on neurofeedback techniques has shown promise in rapid training of new emotional responses [17–19], and has been suggested as a possible treatment for psychopathy [17]. Other possible techniques for influencing choices include Transcranial Magnetic Stimulation, Deep-Brain Stimulation, Transcranial Direct Current Stimulation [20], and Optogenetics, offering the prospect of profound manipulation using genetic manipulation and optic stimulation. These technologies can directly modify behaviours, perhaps even addictive behaviour [21].

The hormone and neurotransmitter oxytocin is a substance with effects on moral behaviour. Oxytocin is naturally elevated by sex and touching. But it can also be elevated by nasal spray. It facilitates birth and breastfeeding in humans and other mammals, but it also appears to mediate maternal care, pair bonding, and other pro-social attitudes, like trust, sympathy and generosity [22]. Kosfeld and collaborators investigated the relationship between oxytocin and trust in a simple game of cooperation [23]. Participants who had been administered oxytocin exhibited significantly more trusting behaviour than those administered the placebo. However, oxytocin's effects on trusting and other pro-social behaviour towards others appears to be sensitive to the group membership of the others [24, 25], suggesting that the pro-social effects of oxytocin may be limited to in-group members. Further experiments by De Dreu's group indicated that oxytocin can also *reduce* pro-social behaviour towards out-group members where this helps one's in-group [24]. Since in-group favouritism seems to drive class and racial discrimination, which in extreme cases manifests itself in genocide and terrorism, administration of oxytocin would not by itself be an effective cure against these evils.

Another neurotransmitter implicated in moral behaviour is serotonin. Selective serotonin reuptake inhibitors (SSRIs) are commonly prescribed for depression, anxiety, and obsessive compulsive disorder. They help govern activities such as eating and sleeping, and sexual activity. SSRIs work by slowing the reabsorption of serotonin, a neurotransmitter crucially involved in mood, thereby making more of it available to stimulate receptors. SSRIs also seem to make subjects more fair-minded and willing to cooperate. Tse and Bond [26] had subjects play the Dictator game – a game in which a dictator decides how a certain sum of money is to be divided between him or her and another participant – and found that subjects administered the SSRI citalopram divided the sum more fairly than controls. Conversely, depletion of precursor of serotonin (tryptophan), which would lead to reduced levels of serotonin, leads to lower rates of cooperation in the Prisoner's dilemma game [27]. Crockett and colleagues [28] found that depletion of tryptophan led to increased rates of rejection of unfair offers relative to controls. This suggests that SSRIs may make subjects easier to exploit by modulating their assessment of what counts as (unacceptably) unfair.

Ritalin is used in children with Attention Deficit Hyperactivity Disorder to improve impulse control and reduce violent aggression. It has been shown in one study to reduce violence by up to 41% in adults [29].

Even from this brief survey, it can be seen that biomedical interventions *can* influence moral decision-making: the possibility of moral enhancement is not mere science fiction. Indeed, many drugs which are already in use have moral effects [30]. However, currently reported effects are usually small, short-lived, and – particularly as demonstrated with with oxytocin – often highly contextual. There is thus still a great amount of work for science to do before we can confidently use biomedical interventions for moral enhancement. This being so, we propose that a new avenue of enquiry – moral artificial intelligence (moral AI) – could be explored by theorists and scientists alongside continuing research into biomedical enhancement.

5 A Case for "Weak" Moral AI

"Pervasive" or "ubiquitous" computing and the more recent concept of "ambient intelligence" all point to a future wherein we will become increasingly integrated with the technologies that we use to obtain, process and act on available data. In particular, ambient intelligence – a human-centric application of artificial intelligence – refers to a system that gathers information form multiple sensors and processes the functional significance of this information in "awareness" of environmental and user context (see [31]). Whilst research on ubiquitous computing and ambient intelligence has thus far sought to envisage how such technology could make the lives of humans easier or more efficient, we suggest that it could also be employed to make the lives of humans more moral. Broadly, artificial intelligence could be developed to address the limitations of human moral design in "stronger" or "weaker" ways. Strong moral AI – unlike the ambient intelligence paradigm – would involve creating enhanced artificial moral agents. These agents would be virtuously superior to us, created to exhibit the best human qualities: they would (at minimum) be altruistic, co-operative and just (fair). They would constantly review how these virtues should be calibrated and deployed from the consequences of their actions.

However, a proposal to create autonomous agents would be met with the usual concerns about their potential for evolution beyond our control or prediction. Moreover, the employment of strong AI to modify human behaviour would likely undermine human freedom, though one of us has argued this might be a price worth paying in certain circumstances (see [32]). Nonetheless, there is likely to be strong resistance to the employment of strong AI to improve ordinary moral behaviour. We therefore argue that the contender for serious consideration is a type of "weak" moral AI that does not involve creating new agents nor undermining human freedom but, instead, gathers, computes and updates data to assist human agents with their moral decision-making. This data will comprise information about the individual agent and his environment, about his moral principles and values and about the common cognitive biases that affect moral decision-making. The moral AI will use this data to alert the agent to potential influences and biases, will suggest strategies for ameliorating these influences and biases, and will advise the agent of particular courses of action at his request.

To illustrate the potential for moral AI (so conceived) to enhance an agent's moral judgements and decisions we describe a prototype moral AI that has both continuous and situation-specific functionality. As a continuous observer of the agent and his environment, the moral AI alerts the agent to features of his own physiology, mental states or environment, excluding the behaviour and dispositions of other agents, that might impair moral judgement and/or behaviour: it is a moral environment monitor. As a continuous observer of the behaviour of the agent, it can alert the agent to any (self-set) moral targets that he has or is likely to miss if he does not act in a particular way: it is a moral organiser. When it comes to particular moral decisions, the moral AI takes the agent through particular moral considerations and makes him aware of situation-specific biases: it is a moral prompter. If requested, the moral AI takes into account the agent's self-identified moral values to suggest a particular course of action: it is a moral advisor. We flesh these four roles out below and then suggests a more controversial fifth function: AI as a protector against others' immorality, dispositions and behaviour.

5.1 Continuous Function One: Moral Environment Monitor

The first function of the moral AI would be to continuously monitor the agent's physiology, mental states and his environment. In so doing, it would be equipped to alert the agent to particular factors that tend to affect moral decision-making and behaviour. Over time, the moral AI would learn which factors particularly affect the specific agent. The moral AI would therefore function as a bio-feedback facility, where the physiological, psychological and environmental data is analysed from the perspective of optimal moral functioning. The sort of data collected would be likely to include (but would be far from limited to) the following.

The *amount of sleep* an agent has had can affect his moral judgements. For example, a study of sleep-deprived US soldiers showed that the officers' ability to conduct mature and principally oriented moral reasoning was severely impaired during partial sleep deprivation compared to the rested state [33]. The moral AI would alert the agent when his level of tiredness is such that his moral reasoning was likely to be impaired.

The *time between meals* can affect an agent's moral judgements. For example, a recent study reported a disturbing correlation between the sentencing decisions of experienced judges and when they take their food breaks [34]. The probability of a favourable ruling drops dramatically as the judge was further away from the preceding food break. The moral AI would alert the agent when the time since he last ate became likely to affect his moral judgments.

Particular *physiological patterns of arousal* can (sometimes via cognition or emotion) affect the way in which an agent interprets social interactions and the way in which he engages with them. For example, Zillmann's (e.g. [35]) studies supporting his "excitation transfer" hypothesis show that sympathetic (nervous system) arousal elicited even by morally irrelevant sources (such as exercise) can be misattributed by the agent as arising from situations involving some provocation. This misattribution of arousal causes the agent to "misread" his aroused state as anger, which motivates aggressive behaviour. It has been further suggested that the arousal effect may persist even after the arousal has dissipated such that the agent may remain potentially aggressive for as long as the self-generated label of "angry" persists. The moral AI would monitor sympathetic nervous system activity so that the agent would be made aware were there to be the potential for excitation transfer to elicit aggressive responses where they would otherwise not be forthcoming. Crucially, studies have shown that when the source of the arousal is made salient to an agent, he is less likely to misattribute it and excitation transfer does not occur (see [36]).

Studies of *ego depletion* have demonstrated that self-control is a limited resource; the more temptation a subject has resisted in the recent past, the more likely they are to give in to a subsequent temptation [37, 38]. The moral AI would monitor the instances of self-control the agent practices, alerting him when he is likely to find it difficult to continue to resist tempting stimuli.

As surveyed above, *levels of hormones and neurotransmitters* have been shown to influence moral behaviour. Testosterone has been shown to make agents more aggressive and more utilitarian in their judgements (see [39]). Serotonin has been shown to be linked to propensity towards co-operation and fair-mindedness [26]. Oxytocin has been shown to increase trust towards in-group members but reduce pro-social behaviour towards out-group members where this helps one's in-group [24]. The moral AI would monitor hormone and neurotransmitter levels alerting the agent to his current levels and their associated affects on moral judgement and behaviour.

Monitoring the agent's *intake and metabolisation of particular psycho-active substances* such as alcohol and various drugs would allow the agent to know when he might be in a state not conducive to making moral decisions or even at risk of losing control over his actions. In addition to measuring the intake and effects of actual substances, it has recently been suggested that artificial intelligence technology could be designed to *detect an agent's developing drug cravings* and – as a multimedia device – intervene, as the cravings develop, to prevent drug use. In service of these functions, the "iHeal" is described as "an innovative constellation of technologies that incorporates artificial intelligence, continuous biophysical monitoring, wireless connectivity, and smartphone computation" [40, p. 1].

Environmental factors such as *hot environmental temperature* have been shown to influence the agent's behaviour [41]. Hot temperatures have been shown to encourage hostile interpretations of ambiguous situations, without the agent being aware of this influence. *Crowdedness and noisiness* have been proposed to have similar effects (see [42–44]). Even environmental factors such as the *tidiness* of one's desk have been shown to affect deliberative capacities and perception of others [45].

5.2 Continuous Function Two: Moral Organiser

The second continuous function of the moral AI would be to assist agents in setting and meeting particular moral goals. An agent might wish to donate a specific amount of money to charity every year or to spend a certain amount of time volunteering for altruistic causes. Another agent might wish to reduce his carbon footprint or to become better at keeping promises. The AI would be aware of opportunities for the agent to meet his goals (for example new charitable organisations or events; alternative travel options), make suggestions about how best to achieve his goals, and alert him when he misses his targets.

Survey research has shown that agents often give less money than they think: 10 per cent of the respondents to the generosity survey reported tithing 10 per cent of their income to charity although their records showed they gave \$200 or less.[1] These findings suggest that, for monetary donating and other altruistic targets, the moral organisation function of the AI could assist agents in doing what they set out to do, where they really want to achieve their goals.

5.3 Situation-Specific Function One: Moral Prompter

The third function of the moral AI would be to serve as a neutral prompt to moral reflection. The agent would indicate a category of moral choice or dilemma facing him and the AI would run through relevant questions to aid the agent in moral deliberation. The questions would be motivated by a variety of ethical considerations drawn from different accounts of what constitutes right action. In being prompted to think more deeply about his decision, its motivation and ramifications, the agent exerts more control over his choice.

We envisage that the moral AI would be developed to utilise a complex categorisation of specific types of moral choice or dilemma ranging from decisions of justice (relating to punishment but also perhaps to employment), to reproductive decisions, to decisions affecting the environment, to parenting decisions, to decisions about dividing time or loyalties between others, and so on. However, to illustrate the function here we adopt the coarse-grained distinction from psychological research on real-life moral choices between antisocial and prosocial moral decisions. Within this research, the term antisocial is used to refer to dilemmas involving "reacting to a transgression (for example, involving violations of rules, laws or fairness) committed by them; dealing with the temptation to meet their own needs or desires, acquire resources, or advance their own gain by violating rules or laws, behaving dishonestly, immorally, or unfairly" [46, p. 166]. The term prosocial is used to refer to dilemmas involving "dealing with two or more people making inconsistent demands on them, with implications for their relationship with each person; deciding whether or not to take responsibility for helping someone important to them" [46, p. 166].

Based on this categorisation, an agent facing certain types of antisocial dilemma might be prompted with the following questions: "what would be the consequences of your act for your self and others?", "is there some less problematic way of meeting your needs or desires?", "would the act involve crossing a line you promised your self or another you wouldn't cross?", "do you think you will feel shame or remorse if you go ahead with the act?", and so on.

An agent facing certain types of prosocial dilemma might be prompted with the following questions: "have you promised more to one party than to the other?", "would one course of action result in more overall benefit than the other?", "are you being influenced by any irrelevant characteristics of the two parties, such as race or gender?", "do you think that if you have the time and capacity to help the person in need you should?", and so on.

[1] See Price and Smith, "Religion and Monetary Donations: We All Give Less Than We Think", cited: `http://generosityresearch.nd.edu/news/33562-the-flesh-is-weak-churchgoers-give-far-less-than-they-think/`, accessed 28th May 2013.

Depending on the category of choice or dilemma, the AI could also prompt the agent to consider particular measures to reduce bias or irrelevant external influence. One example of this would operate in decisions of justice/fairness. It has been shown that gender can have a distorting influence on decision-making: one study showed that the performance of female (but not male) pianists was judged as less good when their gender was "visible" [47]. The AI could prompt agents faced with justice/fairness decisions to mitigate gender biases by making themselves gender-blind where possible.

5.4 Situation-Specific Function Two: Moral Advisor

The second situation-specific function would allow the agent to ask the AI for moral advice about the course of action he should take. Before offering advice in the first instance, the AI would ask the agent to indicate which of a long list of morally significant values or principles he holds and is guided by. Importantly, he is also asked to assign a weight (between 0 and 1) to each value. Thus, an agent who cared very strongly about not harming others (non-malevolence), a bit about not breaking the law (legality) and not at all about protecting the environment (environmental protection) might assign them weights of 1, 0.5 and 0, respectively. We suggest a non-exhaustive list of possible values to include:

- Autonomy (of others – e.g. not being paternalistic)
- Benevolence (helping others)
- Non-malevolence (not harming others)
- Justice/fairness
- Legality
- Environmental protection
- Family/significant relationships
- Fulfilling duties/commitments/promises
- Maximising net utility (making sure overall benefits outweigh overall costs)

For any given scenario, the AI would compute the extent to which the courses of action open to the agent would uphold or compromise these values (fully uphold value = 1; fully compromise value = -1), amplifying or diminishing based on the weight indicated by the agent. The AI would then use these weighed values to suggest the best course of action. The agent would have the opportunity to feedback about whether he took the advice or not and to change the weighing of his values.[2]

The situation-specific advice could be integrated with the continuous environment monitoring and moral organisation. For example, given that high levels of testosterone and sleep deprivation are factors that make individuals more likely to make utilitarian decisions (to over-value maximising net utility), the AI could warn the agent that, for this reason, his current assessment of the advised course

[2] For a comparable suggestion using just three moral values to make decisions in the healthcare setting see [48].

of action may not be consonant with his more enduring moral principles. Related to moral organisation, the AI might suggest that in a particular instance the agent should "over" prioritise a particular value in order to meet his long-term moral goals. For example, if an agent indicated that he wanted to reduce his carbon footprint by a certain amount that year but had been failing to do so, the AI might recommend increasing the weighing of the value of environmental protection in one instance. In this way, the local advice is sensitive to and informed by the agent's global values and goals.

5.5 A Further Function: Protection from the Immoral?

Whilst the function of the moral AI thus far conceived has been to assist the agent in becoming more moral, there is also the possibility that the moral AI could provide a protective function relating to the potential *im*morality of others. Whilst human beings have evolved an impressive capacity to interpret the intentions and emotions of others based on cues such as facial expression and posture, technology is being developed that might soon surpass these capacities. In the past decade, devices such as polygraph lie detectors have provided us with a way of improving our ability to predict when someone is lying. More recently, facial expression recognition technology and other technologies predicting "malintent" have been developed with intended applications in marketing and surveillance.

Within the commercial domain, Affectiva have created Affdex, a product that reads facial expressions to measure the emotional connection people have with advertising, brands and media. According to Affectiva, "Affdex reads emotional states such as surprise, dislike and attention from facial expressions using a webcam. It employs advanced computer vision and machine learning techniques to recognise and automate the analysis of tacit expressions, and it applies scientific methods to interpret viewers' emotional responses quickly and at scale".[3] A comparable technology that could alert an agent to others' micro-expressions relating to, for example, suspicion or insincerity could prevent him from mis-placing his trust.

Going beyond analysis of facial expressions, the U.S. Department of Homeland Security and scientists funded from the Seventh Framework Programme of the European Union have concurrently been developing technologies that can detect "malintent" or "abnormal behaviour and threats", respectively. The U.S. version – Future Attribute Screening Technology (FAST) – uses distant sensors to measure changes to the state of an individual's autonomic nervous system (heart rate, breathing rate, body temperature etc.) when asked questions about his possible malintent.[4] These changes are compared to to the individual's baseline parameters, measured at an earlier point. FAST therefore predicts the probability of malintent based on physiological data. In the EU, scientists have been

[3] See: `http://www.affdex.com/technology/#pane_overview`, accessed 16th June 2013.

[4] See: `http://www.dhs.gov/xlibrary/assets/privacy/privacy_pia_st_fast.pdf`, accessed 16th June 2013.

working on the Automatic Detection of Abnormal Behaviour and Threats in crowded Spaces (ADABTS).[5] The crowded spaces thought to be of particular importance are the airport, the football stadium, and the town center. Within the context of the ADABTS project, three criteria of abnormal behaviour are thought to be relevant: statistical infrequency, violation of norms, and unexpectedness. Abnormal behaviours within these criteria can comprise actions (such as rushing through the crowd or using an emergency exit) or physiological indicators (similarly to FAST) (see [49] for further discussion of these two projects).

An AI that could detect and interpret micro-expressions and even suspicious physiological indicators in others could serve to protect the agent from harm. However, the possibility of this technology raises interesting ethical questions relating to privacy. When in public places, people's faces – and the expressions they make – are usually visible. When having a conversation with someone, I have no grounds for complaint if they pay attention to my face as a source of information about my thoughts and my mood – such attention and interpretation is part-and-parcel of human interaction. However, the potential use of a facial expression recognition technology raises the question whether this capacity can be *too* good. Is there a point at which superior tracking and analysis of what is visible to all somehow invades my privacy? Perhaps, in the same way as I might have legitimate complaint if you were to use a magnifying glass to improve the spatial resolution with which you see my face, I might have legitimate complaint if you use expression recognition technology to improve the temporal resolution.

We might compare the AI function described with a person who has been trained in Ekman and Friesen's facial action coding system (FACS) [50] – a catalogue of 44 facial actions corresponding to independent movements of the 27 facial muscles. Even though this person would be better at reading faces than the average person, there does not seem to be the same worry about privacy. Perhaps the reason is not only to do with computing power/resolution. The worry might instead arise from confusion between prediction and knowledge. We do not worry about the FACS-trained person because we know she is using her skills to make mere predictions about what our faces are revealing. In contrast, the illusion created by the expression recognition technology is that it knows – that it can somehow see inside our minds. But this is not the case. What the technology does is to make predictions based on information about correlations between facial movements and mental/emotional states. It is better at predicting but it still does not know. As long as the agent using the AI also understood that it offered prediction, not knowledge, he could usefully incorporate the information into his calculations about how best to act.

The biosensory technologies used to predict "malintent" equally do not know anything about the agent's mental states – they too work with correlations based on physiological data. However, there is an important difference between these and the facial expression recognition technologies: whilst we know that our faces are visible to our interlocutors, our internal physiology (although sometimes

[5] For an overview, see `http://ec.europa.eu/enterprise/newsroom/_getdocument.cfm?doc_id=6901`, accessed 16th June 2013.

inferable from our external physiology) is not something we usually "reveal" to others. If an agent could measure my heartbeat from a distance, I might feel that my privacy have been somewhat violated. Whilst these ethical questions are up for discussion, the possibility of guarding against immorality would add a desirable fifth function to the moral AI.

6 Preserving Moral Autonomy and Group Level Moral AI

As was noted, work on moral enhancement often comes under pressure to explain how it will be decided that a particular psychological or behavioural effect constitutes *moral* enhancement given that different people have different ideas about what morality demands. A particular strength of the proposed core functionality of moral AI is that it allows agents to decide how much weight they want to give to particular values, thus allowing for multiple moral perspectives.

It might be objected that the agent is *not* completely free to set his own moral standards given that the list suggests the values and principles that an agent might want to assent to. Moreover, the list would not include things like "racial discrimination" or "gratuitously inflicting harm" or "subordinating women". However, whilst it is important to preserve moral pluralism, this is not the same as endorsing complete moral relativism. Common human morality, whilst not always in agreement on finer points, does require some objective standards (see [51]). Gratuitously inflicting harm and the like are "values" that would be immoral on any plausible moral account. We therefore argue that the scheme we propose above preserves enough freedom of values to preserve agents' moral autonomy. Indeed, not only does the proposed moral AI preserve the agent's moral autonomy, it in fact enhances it by prompting the agent to reflect on and assent to moral values and principles, and by equipping the agent to be more successfully guided by the values and principles he endorses. Cognitive biases and other psychological limitations necessarily undermine self-governance. An instrument that could assist an agent in overcoming said biases and limitations would therefore promote the agent's autonomy.

Where questions about imposing moral standards might reoccur would be if moral AI was attempted at the group level. We might imagine that moral AI could be used to advise on the distribution of health resources within the NHS or to direct effective aid programmes. However, the AI would have to direct action based on a particular distributive principles. For example, health resources could be distributed on strict utilitarian principles (according to which the most justified distribution is one that maximises overall utility regardless of the "starting positions" of those who benefit and do not benefit), on egalitarian principals (according to which the most justified distribution is one that leaves all individuals as equal as possible in the number or amount of resources they benefit from), on sufficientarian principles (according to which the most justified distribution is one that results in the greatest number of individuals being above some critical threshold of advantage) or on prioritarian principles (according to which the correct distribution is one that favours directing resources to individuals as a function of degree of disadvantage).Whilst the moral AI will be able to provide the solution based on any of these principals, someone has to decide which principle to follow. It might be argued that when a decision is made on behalf of a group a moral standard is imposed on those not involved in making the decision and those in disagreement with the decision.

However, policy decisions like this have to be made all the time, and there would be no sense or value in resisting a technology that could better follow the principles underlying the policy on the grounds that the particular principles in operation are contestable. If an AI were able to accumulate and analyse relevant knowledge in ways humans could not do, we would have no good reason to not to exploit this knowledge to better meet our collective goals and needs.

7 Conclusion

We conclude by reflecting on a couple of issues that, until something like the moral AI we describe becomes reality, can only be treated speculatively. One interesting question is whether our ideas about moral competence might change if the use of moral AI were to be widespread and effective. There could be a risk that, since (we assume) the use of moral AI technology could help agents be more moral, moral competence might become entangled with technological competence. Even if the technology were to be very easy and intuitive to use – let us imagine even easier than our current smart phones – there would still need to be some attention to and understanding of its functionality.

It is also difficult to predict whether the use of moral AI would make people think more or less about the choices they make, and how this would affect responsibility for these choices. Even though the moral AI we have described involves much agent engagement – it alerts him and he consults it – it is not clear this would necessarily result in deeper reflection on the part of the agent. Especially where the moral AI offers advice on the best course of action, it might be that agents begin to defer to the AI, thus thinking less about their choices and dilemmas. However, whilst this is an empirical question, it seems psychologically unlikely that agents would blindly follow the advice of the AI where doing so repeatedly resulted in their taking courses of action they later regretted. Intuitively, dissatisfaction with the advice would prompt agents to consider why they disagree with the AI and perhaps to them modifying the weight they indicate they want to give to particular values. The agent thus remains self-governing. Where the agent finds himself repeatedly happy with the advice he receives (again, based on his indicated values), far from compromising autonomy, these would be instances of the AI facilitating, even enhancing self-governance.

Finally, it must be remembered that artificial intelligence, even if one day superior to average human intelligence, will never be infallible. Human reflection and judgement should rarely if ever be eliminated from the process of making important moral decisions. Just as we hold people blameworthy for failing to keep a check on their prejudices and biases, individuals using moral AI would also be blameworthy for failing to reflect on whether the suggested course of action was indeed the best. The function of moral AI as we have outlined in this chapter is not for it to supplant human decision-making. Rather, it should serve as an aid to living a morally better life – an aid that has the capacity to obtain and analyse a far greater amount of relevant information than the agent can alone. Given our significantly limited moral psychology, we should welcome the development of any technology which could aid us in this way.

References

1. Douglas, T.: Moral enhancement. Journal of Applied Philosophy 25(3), 228–245 (2008)
2. Douglas, T.: Moral enhancement via direct modulation: A reply to John Harris. Bioethics 27(3), 160–168 (2011)
3. Harris, J.: Moral enhancement and freedom. Bioethics 25(3), 102–111 (2011)
4. Terbeck, S., Kahane, G., McTavish, S., Savulescu, J., Cowen, P.J., Hewstone, M.: Propranolol reduces implicit negative racial bias. Psychopharmacology 222(3), 419–424 (2012)
5. Richerson, P.J., Boyd, R.: Not by Genes Alone: How Culture Transformed Human Evolution. University of Chicago Press (2005)
6. Maynard Smith, J., Szathmáry, E.: The Major Transitions in Evolution. Oxford University Press (1997)
7. Williams, G.C.: Adaptation and natural selection: A critique of some current evolutionary thought. Princeton University Press (1966)
8. Hamilton, W.D.: The evolution of altruistic behaviour. American Naturalist 97(896), 354–356 (2011)
9. Reeder, G.D., Brewer, M.B.: A schematic model of dispositional attribution in interpersonal perception. Psychological Review 86(1), 61–79 (1979)
10. Taylor, S.E.: Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis. Psychological Bulletin 110(1), 67–85 (1991)
11. Alexander, R.D.: The Biology of Moral Systems. De Gruyter, Hawthorne (1987)
12. van Vugt, M., Hart, C.M.: Social identity as social glue: the origins of group loyalty. Journal of Personality and Social Psychology 86(4), 585–598 (2004)
13. Kunda, Z.: The case for motivated reasoning. Psychological Bulletin 108(3), 480–498 (1990)
14. Persson, I., Savulescu, J.: Unfit for the Future? The Need for Moral Enhancement. Oxford University Press (2012)
15. Persson, I., Savulescu, J.: The perils of cognitive enhancement and the urgent imperative to enhance the moral character of humanity. Journal of Applied Philosophy 25(3), 162–177 (2008)
16. Terbeck, S., Kahane, G., McTavish, S., Savulescu, J., Levy, N., Hewstone, M., Cowen, P.J.: Beta adrenergic blockade reduces utilitarian judgement. Biological Psychology 92(2), 323–328 (2013)
17. Sitaram, R., Caria, A., Veit, R., Gaber, T., Rota, G., Kuebler, A., Birbaumer, N.: fMRI brain-computer interface: A tool for neuroscientific research and treatment. Computational Intelligence and Neuroscience, 1:1–1:10 (2007)
18. Sitaram, R., Caria, A., Birbaumer, N.: Hemodynamic brain-computer interfaces for communication and rehabilitation. Neural Networks 22(9), 1320–1328 (2009)

19. Caria, A., Sitaram, R., Veit, R., Begliomini, C., Birbaumer, N.: Volitional control of anterior insula activity modulates the response to aversive stimuli. A real-time functional magnetic resonance imaging study. Biological Psychiatry 68(5), 425–432 (2010)
20. Cohen Kadosh, R., Soskic, S., Iuculano, T., Kanai, R., Walsh, V.: Modulating neuronal activity produces specific and long-lasting changes in numerical competence. Current Biology 20(22), 2016–2020 (2010)
21. Carter, A., Hall, W., Nutt, D.: The treatment of addiction. In: Carter, A., Capps, B., Hall, W. (eds.) Addiction Neurobiology: Ethical and Social Implications, pp. 29–50 (2009)
22. Insel, T.R., Fernald, R.D.: How the brain processes social information: Searching for the social brain. Annual Review of Neuroscience 27, 697–722 (2004)
23. Kosfeld, M., Heinrichs, M., Zak, P., Fischbacher, U., Fehr, E.: Oxytocin increases trust in humans. Nature 435, 673–676 (2005)
24. De Dreu, C.K.W., Greer, L.L., Handgraaf, M.J.J., Shalvi, S., Van Kleef, G.A., Baas, M., Ten Velden, F.S., Van Dijk, E., Feith, S.W.W.: The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. Science 328(5984), 1408–1411 (2010)
25. De Dreu, C.K.W., Greer, L.L., Van Kleef, G.A., Shalvi, S., Handgraaf, M.J.J.: Oxytocin promotes human ethnocentrism. Proceedings of the National Academy of Sciences 108(4), 1262–1266 (2011)
26. Tse, W.S., Bond, A.J.: Serotonergic intervention affects both social dominance and affiliative behaviour. Psychopharmacology 161(3), 324–330 (2002)
27. Wood, R.M., Rilling, J.K., Sanfey, A.G., Bhagwagar, Z., Rogers, R.D.: Effects of tryptophan depletion on the performance of an iterated Prisoner's Dilemma game in healthy adults. Neuropsychopharmacology 31(5), 1075–1084 (2006)
28. Crockett, M.J., Clark, L., Tabibnia, G., Lieberman, M., Robbins, T.W.: Serotonin modulates behavioral reactions to unfairness. Science 320(5884), 1739 (2008)
29. Lichtenstein, P., Halldner, L., Zetterqvist, J., Sjölander, A., Serlachius, E., Fazel, S., Långström, N., Larsson, H.: Medication for attention deficit-hyperactivity disorder and criminality. New England Journal of Medicine 367(21), 2006–2014 (2012)
30. Levy, N., Douglas, T., Kahane, G., Terbeck, S., Cowen, P., Hewstone, M., Savulescu, J.: Are you morally modified? the moral effects of widely used pharmaceuticals. Philosophy, Psychiatry and Psychology (forthcoming)
31. Charalampidou, M., Mouroutsos, S., Pavlidis, G.: Identifying aspects of ambient intelligence through a review of recent developments. Journal of Advanced Computer Science & Technology 1(3), 82–100 (2012)
32. Savulescu, J., Persson, I.: Moral enhancement, freedom and the God machine. The Monist 95(3), 399–421 (2012)
33. Olsen, O.K., Pallesen, S., Eid, J.: The impact of partial sleep deprivation on moral reasoning in military officers. Sleep 33(8), 1086–1090 (2010)
34. Danziger, S., Levav, J., Avnaim-Pesso, L.: Extraneous factors in judicial decisions. Proceedings of the National Academy of Sciences 108(17), 6889–6892 (2011)
35. Zillmann, D.: Cognition-excitation interdependencies in aggressive behavior. Aggressive Behavior 14(1), 51–64 (1988)
36. Reisenzein, R., Gattinger, E.: Salience of arousal as a mediator of misattribution of transferred excitation. Motivation and Emotion 6(4), 315–328 (1982)
37. Baumeister, R.F., Bratslavsky, E., Muraven, M., Tice, D.M.: Ego depletion: Is the active self a limited resource? Journal of Personality and Social Psychology 74(5), 1252–1265 (1998)

38. Baumeister, R.F.: Ego depletion and self-control failure: An energy model of the self's executive function. Self and Identity 1(2), 129–136 (2002)
39. Montoya, E.R., Terburg, D., Bos, P.A., Will, G.J., Buskens, V., Raub, W., van Honk, J.: Testosterone administration modulates moral judgments depending on second-to-fourth digit ratio. Psychoneuroendocrinology (2013)
40. Boyer, E.W., Fletcher, R., Fay, R.J., Smelson, D., Ziedonis, D., Picard, R.W.: Preliminary efforts directed toward the detection of craving of illicit substances: The iHeal project. Journal of Medical Toxicology 8(1), 5–9 (2012)
41. Anderson, C.A., Deuser, W.E., De Neve, K.M.: Hot temperatures, hostile affect, hostile cognition, and arousal: Tests of a general model of affective aggression. Personality and Social Psychology Bulletin 21(5), 434–448 (1995)
42. Berkowitz, L.: Frustration-aggression hypothesis: Examination and reformulation. Psychological Bulletin 106(1), 59–73 (1989)
43. Berkowitz, L.: On the formation and regulation of anger and aggression: A cognitive-neoassociationistic analysis. American Psychologist 45(4), 494–503 (1990)
44. Berkowitz, L.: Pain and aggression: Some findings and implications. Motivation and Emotion 17(3), 277–293 (1993)
45. Sitton, S.: The messy desk effect: How tidiness affects the perception of others. Journal of Psychology: Interdisciplinary and Applied 117(2), 263–267 (1984)
46. Wark, G.R., Krebs, D.L.: Sources of variation in moral judgment: Toward a model of real-life morality. Journal of Adult Development 4(3), 163–178 (1997)
47. Davidson, J.W., Edgar, R.: Gender and race bias in the judgement of Western art music performance. Music Education Research 5(2), 169–181 (2003)
48. Pontier, M.A., Hoorn, J.F.: Toward machines that behave ethically better than humans do. In: Proceedings of of the 34th International Annual Conference of the Cognitive Science Society, Austin, TX, pp. 2198–2203 (2012)
49. Sutrop, M., Laas-Mikko, K.: From identity verification to behavior prediction: Ethical implications of second generation biometrics. Review of Policy Research 29(1), 21–36 (2012)
50. Ekman, P., Friesen, W.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press (1978)
51. Santoni de Sio, F., Maslen, H., Faulmüller, N.: The necessity of objective standards for moral enhancement. AJOB Neuroscience 3(4), 15–16 (2012)

Emotion, Artificial Intelligence, and Ethics*

Kevin LaGrandeur

[1] New York Institute of Technology, NY, USA
[2] Institute for Ethics and Emerging Technologies, Hartford CT, USA
klagrand@nyit.edu

Abstract. The growing body of work in the new field of "affective robotics" involves both theoretical and practical ways to instill – or at least imitate – human emotion in Artificial Intelligence (AI), and also to induce emotions *toward* AI in humans. The aim of this is to guarantee that as AI becomes smarter and more powerful, it will remain tractable and attractive to us. Inducing emotions is important to this effort to create safer and more attractive AI because it is hoped that instantiation of emotions will eventually lead to robots that have moral and ethical codes, making them safer; and also that humans and AI will be able to develop mutual emotional attachments, facilitating the use of robots as human companions and helpers. This paper discusses some of the more significant of these recent efforts and addresses some important ethical questions that arise relative to these endeavors.

Keywords: artificial intelligence, affective robotics, ethics, artificial emotions, empathic AI, artificial conscience.

1 Introduction

Many current ideas about creating emotions in Artificial Intelligence (AI) are highly speculative. They are premised upon a future in which we have sentient AI ("strong" AI), and that is a future that could be quite a long way off – or that may never happen. These ideas include two parallel but separate camps of thinkers: those who discuss "friendly AI" [1, 2] and those who contemplate what are variously called "moral machines", "robot ethics", or "Artificial Moral Agents (AMA)" [3–5]. As opposed to these foci, this chapter will focus on more recent and actual developments regarding the creation of various emotional states in AI, the social motives for doing so, and the ethical dimensions of those efforts and motives.

As a start to this endeavor, we can examine the relatively new field of "affective robotics", the recent attempts to induce an affective state (emotion) in various types of Artificial Intelligence (AI). These efforts stretch back a little more than a decade, and include numerous ways of simulating emotions (and more recently,

* This work was supported by a Grant from the Culture Agency of the European Union, Agreement Number 2013 - 1572 / 001 - 001 CU7 MULT7, Metabody Project; as well as by the New York Institute of Technology (NYIT).

J. Romportl et al. (eds.), *Beyond Artificial Intelligence,*
Topics in Intelligent Engineering and Informatics 9, DOI: 10.1007/978-3-319-09668-1_7

seeking ways to actually instill emotions) in AI and the different motives for doing so. One article that surveys early work on social robots, and which thus cites numerous other early studies, notes, in summary, that there are three main motives for creating artificial emotions:

> [to] facilitate believable human-robot interaction ... [to] provide feedback to the user, such as indicating the robot's internal state, goals, and (to some extent) intentions ... [and to] act as a control mechanism, driving behavior and reflecting how the robot is affected by, and adapts to, different factors over time. [6]

What has changed in the eleven years since this survey was written is that the stakes have risen concerning the use of AI because it has become so widespread and has migrated into sensitive areas, such as the military, domestic companionship, and consumer health care. Accordingly, we will examine AI, emotion, and Human-AI interaction within these particular contexts. There are several motives for trying to create emotions in Artificial Intelligence within these contexts. One is safety, and the other is the attractiveness of AI to humans. If AI continues to become more intelligent, and eventually is able to act autonomously, attractiveness and safety will be paramount, because the biggest sector of robot manufacturing is that for "service robots", which are made primarily for consumers and the military. These kinds of robots – as opposed to "industrial robots" used for manufacturing – include everything from automated milking machines to military drones (which at the moment are the two most numerous types of service robots); this category also includes domestic and personal robots, such as robotic vacuums and "companion" robots of the type used, for instance, as health care aids for the elderly and for autism therapy.

If we look at the statistics given by The Institute of Electrical and Electronics Engineers (IEEE) concerning "World Robot Population", it is clear just how many more service robots exist compared to industrial ones, and how rapidly their numbers are increasing. According to these statistics, by 2011, the total number of robots in use worldwide was 18.2 million: only 1.2 million of those were industrial, and the rest – 17 million – were service robots [7]. Another way of looking at this is via the statistics of the International Federation of Robotics (IFR), the major trade group for robot manufacturers. Their data show 2.5 million personal and domestic robots sold in 2011 alone [8, p. 15]; that is more than all of the industrial robots ever sold, from the 1960's to 2011, which amounts to about 2.3 million [8, p. 10].

2 Programming Emotion in an Effort to Make AI Safe, Friendly, and Attractive

2.1 Military Robots and Emotion: Seeking a Path to Safety

With such rapidly increasing numbers and, obviously, increasing use of machines that can be very powerful, safety is an important issue. And because of their numbers and social impact, the two types of AI that are receiving the most funding and attention for developing affects or emotions are the personal robots I mention above and military robots. Given that safety is the key factor that unites both of these types of AI, we will focus first on that factor, and first on military AI, because that is where this issue of safety is the supreme consideration. Indeed, if AI continues to become more intelligent and, especially, more autonomous, safety will become an ever more pressing issue in military robotics – which is presently the largest sector of the robot industry. Making computerized, autonomous weapons is clearly fraught with concerns such as distinguishing between civilians and soldiers, and between friendly soldiers and the enemy, among other things. Furthermore, it is clear that making military robots more autonomous is exactly what the military aims to do. A relatively recent Call for Proposals by the United States Army makes this clear:

> Armed UMS [Unmanned Systems] are beginning to be fielded in the current battlespace, and will be extremely common in the Future Force Battlespace ... This will lead directly to the need for the systems to be able to operate autonomously for extended periods, and also to be able to collaboratively engage hostile targets within specified rules of engagement ... with final decision on target engagement being left to the human operator ... *Fully autonomous engagement without human intervention should also be considered.* [9, italics added for emphasis]

Ronald Arkin cites this passage in his work, and then points out that there were, as of 2009, already a number of semi-autonomous intelligent weapons in use by the United States Armed Forces and by other forces. One prime example, which is actually fairly old, is the Cruise Missile, which once launched does most of the work of acquiring and destroying its target. Another, more chilling example is a South Korean intelligent weapons platform that can "detect and identify targets in daylight within a 4 km radius, or at night using infrared sensors within a range of 2 km, providing for either an autonomous lethal or non-lethal response. Although a designer of the system states that 'the ultimate decision about shooting should be made by a human, not the robot,' the system does have an automatic mode in which it is capable of making the decision on its own" [10, pp. 4–6].

The reason Arkin cites the Army's Call for Proposals and then the examples mentioned above is in order to give a rationale for his own project, which is an attempt to create the software architecture for an artificial conscience that would serve as an ethical governor of autonomous smart weapons. The need for this is, as noted above, very clear. There have already been some documented problems with smart weapons being able to distinguish the proper "targets", such as the infamous malfunction of a smart antiaircraft gun in South Africa, when it turned and began killing the soldiers operating it [11]. The main basis for Arkin's artificial conscience would be a reason-based decision tree, but he admits that this plan is not completely sufficient. He points out that the use of emotions may be needed as part of the ethical apparatus. In addition to the rational decision tree based primarily on Kantian deontology (that is, an ethical system focused on sense of duty) combined with the Army's Codified Laws of War, he concedes that there should also probably be some sort of ethical adaptor based on "artificial affective function (e.g., guilt, remorse, grief)" that would motivate a weaponized AI to review and correct any mistakes it had made in using lethal force [10, pp. 20–21]. This would be done especially focusing on guilt. The coding that would induce robotic "guilt" would be the robot's own monitoring of certain measurable parameters, such as "noncombatant casualties and damage to civilian property, among others", or "criticism" from human monitors [10, p. 74]. Formally, this affective function would be expressed like this:

$$\text{IF Vguilt} > \text{Maxguilt THEN Pl-ethical} = \emptyset$$

"where Vguilt represents the current scalar value of the affective state of Guilt, and Maxguilt is a threshold constant ..." If the threshold constant is exceeded, then ethics have been breached, and the weapon is automatically disabled [10, p. 74]. The biggest ethical problem with this idea is that such affective functioning only occurs after some kind of heinous humanitarian violation has occurred. The other problem is that these robot "emotions" are only vague simulations – or not even that, but just "diagnostic troubleshooting", as is done now with malfunctioning computers, but under a different name. Another problem is that the emotions Arkin wants to use in military robots are still based on cold calculation of assessment criteria, not on empathy, sympathy, or, as Arkin himself admits, compassion [10, p. 75]. Thus, calling this development "affective" computing or considering it emotion-based is inaccurate.

We should note that Arkin's stated goal is not to help produce better weapons but to prevent what he sees as the inevitable weaponization of AI by the military from being an unharnessed free-for-all, with huge inadvertent slaughters of innocent non-combatants. This goal of preventing unharnessed slaughter by military robots is a noble one, but, as can be seen above, the coding is just not complex enough, nor the AI advanced enough, to instill the necessary emotions and behavior desired. This is a practical engineering problem that is widespread right now, though small steps toward creating at least simple simulations of some emotions in limited contexts are appearing. Chapter 12 by Alidoust and Rouhani in this present volume is an example of that. They present a model for simulation of four emotions (anger, happiness, nervousness, and relief) which, though too simple to imitate human behavior (these behaviors have very narrow determinants in the modeling agents), is a step toward investigating more complex behavior containing more variables and complexities.

However, as Hamid Ekbia argues in his Chapter 5 of this volume, attempts to instantiate emotions in AI may always be doomed to simplistic imitations because of an error in basic assumptions about how emotions work in humans. That is, as he maintains, the approach of AI researchers has to this point been based on what he calls a monadic rather than a dyadic model of emotions. The former model is based on the idea that emotion is internally generated, whereas the latter model, which he argues is more accurate, is based on the idea that emotions are dynamic, relational, and intersubjective – they are built on changing relationships with the external world and its inhabitants, and so they cannot simply be internally generated in robots by a program. I agree. Models like Arkin's and the model referred to by Alidoust and Rouhani (the CCC model), are too focused on autonomous action, as opposed to dynamic interaction, to be a good source of complex, human-like emotion. Beyond these practical problems, there are other, more abstract philosophical considerations, which we will examine later in the final section of this paper.

2.2 Companion Robots and Emotion: Not Just Safe, but Also Attractive AI

Rossler's Benevolent AI as a Combined Attempt at Safe Military and Personal Robots

Arkin's ideas and plans are meant for relatively near-term deployment, but they also assume that, in addition to increased autonomy, robots and other military AI will continue to become more intelligent than they are now. Other theories for developing emotional AI in order to protect their human creators assume much more intelligence, including sentience – a concept known as "strong AI" and, as I mentioned in my introduction, most computer scientists consider this type of AI a long way off, if it is possible at all. But there is at least one other theorist who, like Arkin, is trying to make AI friendlier in the near term. That is the German physicist and complexity theorist Otto Rossler. His ideas are laid out in a number of articles, but the most important one is his 2004 article, "Nonlinear Dynamics, Artificial Cognition and Galactic Export" [12]. He claims that, by way of his own mathematical models regarding what he calls "spatial Darwinism", combined with the type of social bonding (called "imprinting") observed by Conrad Lorenz in his famous twentieth-century experiments with geese, a form of benevolent bonding could be programmed into a machine.

Rossler's theory is complex, but this is the essence of it: first, there is Rossler's concept of spatial Darwinism, which he invented to describe how living things survive, not as a species over time (which is Darwin's theory), but as individuals in one lifetime. He maintains that in order to do this, any living animal needs to adapt constantly by moving an appropriate distance through space at the appropriate time, in order to find necessities like food or a mate, and that the valences for this also include important individuals, such as parent figures. Given this definition, Rossler sees benevolence working is as a subset of the concept of "bonding" outlined by Konrad Lorenz and others as the catalyst for benevolence between animal brains. Because bonding works as an adaptive survival trait, it is, he claims, "programmed" into the neural networks of animals. Therefore, for the same reasons, and by way of the mathematical models regarding "spatial Darwinism", bonding could be programmed into a machine.

The way this would work is that algorithms would command a robot's "autonomous path optimization", which Rossler sees as analogous to human emotion in the way that it works to satisfy its needs. In other words, Rossler sees emotion as a function of primordial drives, and as necessary adaptations for satisfying those drives. Programming a machine to stay in close proximity to a human is thus relatively straightforward, if bonding is related to basic drives and if it demands that a particular human be seen as integral to its survival and valuable in its own right – what Lorenz called "the animal with home-valence", or more simply, a mother figure [12, p. 59]. So the autonomous path optimization algorithms would be set to identify a particular human as the "animal with home valence", the equivalent of the mother, for most animals. As a result, the machine would become automatically socially bonded to that human upon first viewing him or her. Then, whenever the human shares things with the robot, as he or she would with a human child, the bonded robot, like a child, would learn to share in return, triggering a learning experience that would initiate an evolving, recursive loop of benevolence between it and humans. A practical problem with this theory is that it depends on the human feeling attachment to the machine, which apparently would be instigated by the robot's following the human around loyally, like a baby goose. Rossler assumes this would please the human, not annoy him. Moreover, a big philosophical problem here is that this theoretical architecture collapses the difference between emotional bonding and simple proximity. And can emotion really be equated to simple "path optimization" for one's survival needs, as Rossler posits, and therefore replicated by a simple algorithm? Although Rossler's model depends more than Arkin's does on "dyadic" relationships to form emotions, as Ekbia and I agree would need to happen, this relational affective model is overshadowed by the fact that the main theory behind Rossler's concept is still "monadic", for the most part (I use Ekbia's apt terms here). It is based on basic drives as the sole reason for emotion, which is too reductive. As Ekbia notes in Chapter 5:

> According to [psychologist Sylvan] Tomkins, our behaviors are largely regulated by affects, which are sustained and general in character, as opposed to drives, which are spatially and temporally specific and hence weak in motivating behavior. Affects, as such, take priority over drives. The hunger drive, foundational to behaviorism and also to Freud's theory of sexuality, for instance, is not powerful by itself. It becomes urgent (and so able to compel behavior) when it is amplified by, say, distress or enjoyment. It can similarly be attenuated or blocked by disgust or fear.

In short, as with Arkin's idea, Rossler's is noble in concept because it attempts to keep humans safe and happy – and it implicitly keeps an intelligent and perhaps even sentient robot or AI happy, as well, but it has dubious underpinnings. It ignores more subtle needs met and produced by emotions. How, for instance, does empathy fit in here? Or sympathy, or compassion? These emotions are complex and dependent upon inter-relationships, and they are arguably just as important to keeping humans safe and happy as the sort of harmlessness and loyalty that Rossler has named "benevolence". Given the complexities of true benevolence – or of any other true emotionally-based moral behavior – Rossler's prescription may be one for creating mere "clinginess", as opposed to benevolence, a physical behavior, rather than a true emotional or ethical state.

Benevolence is not merely a behavior, though it is manifested that way, it is a complex ethical stance, a conscious decision, based on a constellation of emotions, experience, and reason, to act for the benefit of another. As such, it entails more than just a simple reward system: a baby (or robot) may instinctively bond to the mother figure (her face and smile), and sharing behavior may be a first lesson in the mutual benefit of cooperative social behavior, but is it any more than that? From that step to benevolence also involves things like altruism, empathy and sympathy, and feelings of responsibility. Some of these are mysteriously complex, like altruism – which research indicates may have not only a genetic component, because it is an adaptive trait for preserving the species, but also a strong learned one.

Likewise empathy seems to be an inborn potentiality that needs experiential help and human instruction to develop. It is a mixture of brain maturation (a physical development of the human organism) and experience. One has to experience pain, for example, in order to understand the pain of others; and not only that, but one has to experience that pain in different contexts to fully understand others' pain. Instruction also plays an important part – parents saying to the child, "How would you like it if someone did that to you?" This aspect of gaining experience via necessary pain also poses a moral conundrum concerning the implication that we might then need to make a sentient Artificial Intelligence that could experience pain: is that moral?

Empathy and Attractive Personal Robots

When contemplating domestic robots, such as robotic helpers and companions, there is more to consider than just safety. Rossler's ideas for benevolent robots hint at this. Personal robots – which can be divided into the categories of domestic and companion robots – need to be attractive to consumers, as well as safe for them to use. As discussed in the introduction to this paper, a key to attractiveness – to "believable human-robot interaction" – is that robots need to exhibit emotion in order to cause humans to develop a real bond with them. As we have seen, affective robotics is in a nascent stage, but researchers have found that at least one category of emotion – empathy – is providing a foothold to creating real bonds between humans and AI. This is actually a two-way process. Robots need to exhibit empathy, and they also need to inspire empathy for themselves in humans; in other words, robots need to enable humans to imagine themselves as the robot – which means humanizing the robot in their minds.

Unsurprisingly, some of this has to do with constructing anthropomorphic facial expressions, speech, and gestures, as with Cynthia Breazeal's experiments at the MIT Media Lab in the early 2000's [13–16]. The animatronic robots created at the Media Lab, such as Kismet, Leonardo, and Huggable, which can still be viewed at the Media Lab's website, were built with special emphasis on facial expression, gestures, and reactivity of both to human interaction [17]. Research has indicated that when anthropomorphic robots mirror the facial expressions and body movements of the human with whom they are interacting, it encourages the human to develop empathy with them [18–20]. But, perhaps more surprisingly, perception of empathy – and human empathic responses to robots – can also have to do with the robot's actions, or the functions it performs.

Some recent examples will help illustrate these somewhat different effects, and also what can now be done in this area of affective computing, and what yet remains. The first example demonstrates a practical application for the descendants of MIT Media Lab's experimental creatures, with their anthropomorphic expressiveness. And the examples after that one, the instances cited by Matthias Scheutz, exhibit empathy induced by human response to robot functionality. In 2011, a group of researchers funded by the European Union's Platform Seven Agency used empathic robots as teaching tools for elementary school students. As they said, "The goal of LIREC [Living with Robots and Interactive Companions project] was not to build robot companions that replace human contact, but rather to design companions that fulfill their tasks and interact with people in a socially and emotionally acceptable manner" [21, p. 1]. In one of their experiments, reported in a recent article, they used an empathic robot called iCat to teach students to play chess [22]. This robot, made by a Dutch company, and which one can see in the article referenced above [21], looks like a small, plastic cat. It is yellow, is in a sitting position, has tactile sensors in its head and front paws so that it can tell when it is being touched and can react to that. It also has auditory sensors embedded in part of its anatomy, and a tiny webcam mounted in its nose. Most importantly, it has a mobile set of facial characteristics: its mouth, eyes, and eyebrows all move in numerous ways so that, like the MIT creations, it can exhibit facial expressions. Its programming allows it to react to the movements and statements of its human partner. The ability to read human facial expressions is provided by a special software program that also enables it to operate a set of six "model" emotional faces in response to human interaction. This facial expression and recognition software, interestingly, was developed as a (successful) experimental therapy to teach autistic children to better read non-verbal cues [23]. When the students learning to play chess had trouble, the robot would use one of four empathic responses: encouraging comments, offering help, making a bad move intentionally, or scaffolding (which they defined as "providing feedback on the user's last move and, if the move is not good, let the user play again") [22, p. 3].

Such experiments as this show that some forward progress is being made in practical applications of empathic robotics, but these successes should not be overestimated. Concerning the more complex artificial emotions of the type Arkin and Rossler want to achieve, AI is not powerful enough yet to support this intricate function of sentience. Furthermore, even in the applications discussed above, robots do not really feel empathy. As of now, they just simulate the physical markers of it. But efforts to create the appearance of empathy in robots, the physical markers, have been pretty successful. Consequently, inducing empathy *in humans* toward robots has indeed met with success. Studies show that humans develop real attachment and empathy toward robots. One of the earliest experiments to show this was done by Freedom Baird at MIT's Media Lab in 1999 [24]. Baird was taking care of two gerbils and a simple social robot called a "Furby" to see how the two compared. She noticed that neither the gerbils nor the Furby liked to be held upside down: the gerbils started struggling after about eight seconds of being held that way, and the Furby was programmed to say, over and over and in a pathetic voice, that it was “scared” when it was held this way. Both exhibits of discomfort – from the gerbil and from the Furby – bothered her. So she gave the same experience to a group of children, and she found that the children reacted empathically to both the gerbils and the Furby, as she did. Children, on average, would turn the gerbil rightside up again after eight seconds, and within a minute would also feel compelled to relieve the “suffering” of the Furby robot).

Now, this shows two interesting things: first, that even though people knew that this Furby was just a robot, they felt compelled to respond to its (artificial) emotions; and second, they responded to it more slowly than they did to a genuine animal. These same results have since been replicated in other experiments: humans respond empathically to robots, as such, but not as readily as to humans or animals [25]. But the fact this empathic response is a one-way phenomenon – humans already respond empathically to robots’ simulated emotions, and also, as we shall see below, to their actions – is a troubling development, ethically.

This phenomenon of unilateral human empathy toward and attachment to robots is discussed by Matthias Scheutz in a recent book chapter related to robot ethics [26]. He gives a lot of examples of this phenomenon gleaned from various studies, and many of them, as noted above, have to do more with functionality rather than anthropomorphic appearances; one that is somewhat surprising to me is the fact that many people form emotional attachments with their robotic vacuum cleaners, called Roombas. These are simple, disk-shaped devices that merely clean one’s floors – one programs them for the time of day they should run, and then they turn themselves on at the designated time and run in a grid-like pattern, bumping into things until they’ve covered the whole room; then they dock themselves to recharge.

Studies cited by Scheutz show that many people personify these simple robots, and some even form a such sense of gratitude toward them that they actually clean the floor themselves in order to give the Roomba “a day off”. Many people also dress them up in costumes that can be bought online that are tailored to fit the robot. Scheutz is very concerned about the possibly dangerous behavior that such attachment could cause. He chiefly worries that such one-way attachments will make humans emotionally vulnerable to manipulation by robots, via their human or corporate makers. For instance, corporations that know their robots are seductive could program them to suggest to their smitten human owners to buy more of the corporation’s products, or to take other actions not necessarily to their benefit.

An even more direct danger comes from soldiers’ emotional attachment to military service robots, such as the bomb-disposal robots used in Afghanistan and Iraq. Scheutz discusses studies that show soldiers can become very devoted to these robots [26, pp. 211–212]. In these cases, it is not just a matter of wanting to give the robot a “day off”, or wanting to dress it up because one is emotionally attached to it: personifying bomb-disposal robots makes soldiers reluctant to trade them in for new ones once they have become too damaged to function properly. Obviously, that could cost them their lives.

3 The Road to Future Developments in Artificial Emotion

This current state of uneven reciprocation of emotions between robots and humans raises some problems, as we have seen, in great part because robots cannot feel emotion or empathy now. The state of computing is just not powerful enough to provide strong AI, and it is not likely to be unless experiments with quantum computing or molecular computing succeed. However, although we may have a long way to go before we can create molecular or quantum-level computing that leads to super-powerful AI, some elements that are part of emotional response in AI are possible now, because of recent, incremental successes replicating cognitive features that contribute to emotions.

For example, Kim and Lipson did an experiment reported in a recent article (2009) on the efficacy of programs that give robots a basic Theory of Mind (ToM) [27]. Essentially, ToM is the ability to understand another's intentions. Humans commonly use ToM to make inferences about others' feelings and states of mind. These investigators created an evolutionary algorithm that allows one robot to infer from another robot's actions what it might do next and how it reasons. Their experiment's main goal was to develop "... controller inference algorithms in robots [that could] help in interaction with non-robotic actors such as humans ..." [27, p. 2072]. The experimental set-up provided one robot whose mission was to find a path across a room to a light source. That path varied continuously, based on the position of the light and other factors. Ultimately, the experiment was successful: because of algorithms that could evolve with experience, one robot was able to continually improve its inferences about what another was going to do. This was in a tightly controlled situation, but the long-range implications are obvious: Kim and Lipson's success in creating ToM in a robot is a small step toward enabling robots and other AI to read the internal state and intentions of humans, and thus to bring them one step closer to a mutual emotional interface.

Most remarkably, there are Theodore Berger's successes with long-term memory re-generation (and generation) by using implanted chips to replace damaged parts of the hippocampus in rats and monkeys [28, 29]. Berger and his team at University of Southern California have succeeded in recording and transforming into computer code long-term memories that are stored in the hippocampus of these animals. In the case of the rats, they had them perform a memory task. Then, they downloaded and transformed the memory of that task into digital code. Afterwards, they removed the section of the rat's hippocampus that carried these memories and replaced that bit of the brain with a special computer chip, onto which they reloaded the artificially stored memories. They found that these rats' memories could be fully restored using this technique. Even more significant was the fact that Berger's team could also generate or enhance memories that had never existed in the animals – for instance, memory of a task that a rat had never done. They were later able to replicate these same results with monkeys [29]. The implications of this are enormous: if memories can be artificially generated, then it brings up the possibility of generating emotion via chips, too, by using them to replace parts of other brain structures, such as the amygdala, where empathy resides. Obviously, there is also an enormous ethical problem here in giving ourselves the ability to generate false memories, or to enhance long-term memories, which could open up many modes of abuse.

4 Conclusion: Further Ethical Considerations

The ethical concerns of Scheutz's, and those of mine that I've discussed to this point, are specific to particular experiments or projects. What about the ethics and philosophical dimensions of the larger project of generating emotions between humans and Artificial Intelligence? First, although the big problem for the more advanced types of projects like Rossler's and Arkin's is our currently insufficient engineering capabilities, there is also a larger philosophical problem: our perennial disagreement as to what basis we should use to define "proper ethics" when discussing and defining values like "benevolence" or "conscience" – especially given different cultural viewpoints. Kantian deontology (based on pure duty or rule-based ethics), Buddhism (based on selfless compassion), and Utilitarianism (based the greater good of the many) are just a few of the philosophical systems that have been proposed as a basis for AI ethics.

Second, if inducing emotions in AI is important to the effort to create "friendly AI" because it is hoped that AI and humans will develop mutual emotional attachments, then the current experiments are working badly, because so far, the emotional attachment has been a one-way occurrence, as Scheutz reminds us. This is potentially problematic for the reasons noted regarding human vulnerability to emotional manipulation. And third, there are the potential philosophical problems we may create for our treatment of a new species, if we ever manage to create sentient, feeling AI. The problems in this scenario are many, but the chief one that concerns us, regarding emotions, is this: As James Hughes points out [30], and as I mentioned earlier in this paper, in order to feel emotions like guilt, compassion, and empathy, we would have to create suffering beings, because only by suffering do we learn to understand others' pain. But creating a suffering being is of dubious morality. So, would these requirements for instilling compassion in an AI be inhumane?

Research on affective robotics raises some other important philosophical questions relative to it, as well as to human progress in the digital age: Do efforts such as the ones I've outlined risk reducing the complexities of emotive human "movement and the non-verbal spectrum to patterns of imitation and functionality", as some have worried [31]? Clearly they do now. But could these theories and programs do more? Perhaps. Anyone who reads the literature can see that the intent of scientists and others involved in this project is clearly to do more than reduce human emotion to imitations and functionalities, but there is no way yet to do that, and they see imitation as an important initial step.

So, does the increasing juxtaposition of the human with the digital undermine the uniqueness and importance of human kinaesthetic communication processes? Right now, yes, but in the future, the answer to this depends on what sort of perspective one takes on the long-term goals of AI researchers and roboticists. From their perspective, they are trying to replicate those same kinaesthetic communication processes, and in all of their spontaneity, because their ultimate goal is to create robots and AI that are humanoid and – importantly – are emergent: that can, in other words, evolve. If that occurs, then perhaps new, hybrid AI-Human kinaesthetic processes will evolve, as well, and that sort of spontaneous, random change would create its own sort of hybridized kinaesthetic dynamic.

References

1. Yudkowsky, E.: Creating friendly AI 1.0: The analysis and design of benevolent goal architectures (2001), `http://intelligence.org/files/CFAI.pdf`
2. Muehlhauser, L., Helm, L.: The singularity and machine ethics. In: Eden, A.H., Moor, J.H., Soraker, J.H., Steinhart, E. (eds.) Singularity Hypotheses. The Frontiers Collection, pp. 101–126. Springer, Heidelberg (2012)
3. Wendell Wallach, C.A.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press (2009)
4. Anderson, M., Anderson, S.L. (eds.): Machine Ethics. Cambridge University Press (2011)
5. Lin, P., Abney, K., Bekey, G.A. (eds.): Robot Ethics: The Ethical and Social Implications of Robotics. The MIT Press (2012)
6. Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots: Concepts, design, and applications. Technical report, The Robotics Institute, Carnegie Mellon University (2002)
7. Guizzo, E.: 6.5 million robots now inhabit the earth. IEEE Spectrum (2008)
8. International Federation of Robotics Statistical Department: World robotics–industrial robots 2012: Executive summary (2012), `http://www.worldrobotics.org/uploads/media/Executive_Summary_WR_2012.pdf`
9. U.S. Army SBIR Solicitation 07.2, Topic A07-032: Multi-agent based small unit effects planning and collaborative engagement with unmanned systems (2007), `http://www.sbir.gov/sbirsearch/detail/212294`
10. Arkin, R.C.: Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. Technical report, Georgia Institute of Technology (2007)
11. Shachtman, N.: Robot cannon kills 9, wounds 14. Wired Magazine (2007)
12. Rossler, O.E.: Nonlinear dynamics, artificial cognition and galactic export. In: Sixth International Conference on Computing Anticipatory Systems, CASYS 2003, pp. 47–67 (2003)
13. Breazeal, C., Scassellati, B.: Infant-like social interactions between a robot and a human caregiver. Adaptive Behaviour 8(1), 49–74 (2000)
14. Breazeal, C., Scassellati, B.: Robots that imitate humans. Trends in Cognitive Sciences 6(11), 481–487 (2002)
15. Breazeal, C.: Emotive qualities in lip-synchronized robot speech. Advanced Robotics 17(2), 97–113 (2003)

16. Breazeal, C.: Designing Sociable Robots. The MIT Press, Cambridge (2002)
17. MIT media lab personal robots group, `http://robotic.media.mit.edu/projects/projects.html`
18. Gazzola, V., Rizzolatti, G., Wicker, B., Keysers, C.: The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. NeuroImage 35(4), 1674–1684 (2007)
19. Oberman, L.M., McCleery, J.P., Ramachandran, V.S., Pineda, J.A.: EEG evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots. Neurocomputing 70(13-15), 2194–2203 (2007)
20. Breazeal, C., Buchsbaum, D., Gray, J., Gatenby, D., Blumberg, B.: Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots. Artificial Life 11(1-2), 31–62 (2005)
21. Castellano, G., Leite, I., Paiva, A., McOwan, P.W.: Affective teaching: learning more effectively from empathic robots. Awareness magazine: Self-Awareness in Autonomic Systems (2012)
22. Leite, I., Castellano, G., Pereira, A., Martinho, C., Paiva, A.: Modelling empathic behaviour in a robotic game companion for children: An ethnographic study in real-world settings. In: Proceedings of ACM/IEEE International Conference on Human-Robot Interaction, HRI 2012, pp. 367–374. ACM (2012)
23. Kaliouby, R.E., Robinson, P.: Mind reading machines: Automated inference of cognitive mental states from video. In: Proceedings of 2004 IEEE International Conference on Systems, Man and Cybernetics, pp. 682–688. IEEE (2004)
24. Glass, I.: Furrbidden knowledge (2011), `http://www.radiolab.org/2011/may/31/furbidden-knowledge/`
25. Humans feel empathy for robots: fMRI scans show similar brain function when robots are treated the same as humans. Science Daily (2011), `http://www.sciencedaily.com/releases/2013/04/130423091111.htm`
26. Scheutz, M.: The inherent dangers of unidirectional emotional bonds between humans and social robots. In: Lin, P., Bekey, G., Abney, K. (eds.) Anthology on Robo-Ethics, pp. 205–221 (2012)
27. Kim, K.J., Lipson, H.: Towards a "theory of mind" in simulated robots. In: Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference, pp. 2071–2076. ACM, New York (2009)
28. Berger, T.W., Hampson, R.E., Song, D., Goonawardena, A., Marmarelis, V.Z., Deadwyler, S.A.: A cortical neural prosthesis for restoring and enhancing memory. Journal of Neural Engineering 8(4), 046017 (2011)
29. Hampson, R.E., Gerhardt, G.A., Marmarelis, V.Z., Song, D., Opris, I., Santos, L., Berger, T.W., Deadwyler, S.A.: Facilitation and restoration of cognitive function in primate prefrontal cortex by a neuroprosthesis that utilizes minicolumn-specific neural firing. Journal of Neural Engineering 9(5), 056012 (2012)
30. Hughes, J.: Compassionate AI and selfless robots: A buddhist approach. In: Lin, P., Abney, K., Bekey, G.A. (eds.) Robot Ethics: The Ethical and Social Implications of Robotics, pp. 69–83 (2012)
31. del Val, J.: METABODY - media embodiment tékhne and bridges of diversity (2012), `http://www.metabody.eu/`

The Stuff That Dreams Are Made of: AI in Contemporary 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* by Ian McDonald, in order to ask about the elements of the Western world view, or cultural imaginarium, that surround the contemporary notion of a strong AI. Drawing from phenomenology and combined with cultural analysis, this paper focuses on the way of depicting AIs, and the notions that presuppose this depiction. These notions are, roughly, reducible to independence and unexpectedness, political awkwardness, openness to the alien and the occidental value of authenticity.

Keywords: artificial intelligence, science fiction, cultural analysis, values, science fiction theory.

1 Introduction

Ian MacDonald's *River of Gods* is a 2004 science fiction novel, a winner of British Science Fiction Award and an object of many affirmative reviews. Moreover, it is one pointed at as an example of the Western author doing its research and presenting a non-Western country, problems of modern Indian socioeconomics as well as culture – Indian Times calling his work "not bad for a firang [foreigner] who has oodles of imagination and chutzpah" [1]. Neither of these assets is a sufficient reason to make this book an object of "thick" [2], detailed description and analysis; nor is the topic of AI itself, common in science fiction and absolutely central for the cyberpunk subgenre, on which motives *River of Gods* is built. The reason for the following close reading is novelty – not of the AI itself, but of the circumstances in which it emerges.

Firstly, it's an AI that escapes out of its traditional cyberpunk birthplace – the Global North (or, more precisely here, the Global North-West), with two implications. It refreshes the "low life, high tech" [3] world and restructures the ways in which AI is perceived and meanings, that are built around it. The novel is also partly biopunk (i.e. dealing with the implications of the prophesied biotechnological revolution), which further ensures, that the text is a part of the most actual version of the cultural imaginarium that surrounds AI. Moving cyberpunk into the Global South-East also creates a critical context towards the Western practices, portrayed as ultimately economy-based. But, I will try

J. Romportl et al. (eds.), *Beyond Artificial Intelligence,*
Topics in Intelligent Engineering and Informatics 9, DOI: 10.1007/978-3-319-09668-1_8

to argue, the text also reveals a value-oriented vision of the consequences of AI emergence; this vision returns the moral center – and the monopoly to deal with AIs – to the West.

The following paper discusses *River of Gods*, a text that is neither utopian nor dystopian. The object of the study will be to reconstruct a vision of the future, not focusing on technical extrapolation, but rather on the social, political and cultural worldview surrounding the AI. Before we turn to the analysis though, certain elements of the author's theoretical standpoint 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 [4], 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. [5]

To acknowledge this claim is to agree that the cultural analysis of strong AI in science fiction is the analysis of now, of today's values and ideas and beliefs that presuppose the fears and hopes. Therefore, 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 [6], 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 [7]. And thus, Husserl's epoché is understood here as temporary refraining. It is refraining from reading the text through the context of SF historians and theorists – or AI theorists, for that matter – in order to let it unfold its own meanings

On the other hand though, mistrust for the author is also needed. Le Guin, in the short text that is both beautiful and insightful, follows up on the mistrust towards extrapolation quoted before: "Prediction is the business of prophets, clairvoyants, and futurologists. It is not the business of novelists. A novelist's business is lying" [5]. The futures envisioned SF are lies – and serve the truth as well, the author says – because they are metaphors. Metaphors for what? Le Guin doesn't answer; nonetheless I would venture an answer: fictional visions of the future are ethical as well as ontological visions. Or, to be more specific, they are metaphors for worldviews (values along with fundamental concepts, often hierarchical), which tie together technology and otherness with ethics.

Therefore, along with reconstructing the *intentio auctoris*, *intentio operis* must also be re-created, from between the lines and sometimes against author's postulates, since only together they fully answer a question of the worldview of the given author – and, to a degree, the culture(s) he lives in. Quoting a Polish philosopher Ryszard Żarowski, 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" [8].

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, one hundred years after India gained its independence. India, divided into quarreling states and faced with a severe drought, fights for water, as well as American favor.

Technologically, the setting of the story follows the development of AI, 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 [9, loc. 150], along with an appropriate level of self-awareness. Generation 2.5 is an AI generally unrecognizable from humans [9, loc. 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 another important part of the story's setting – and it is so for at least two reasons. Firstly, because it's outside the Global North, or "the West"; it's 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, self-emergent and 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 the use of religion in the presented world – and its connection with the concept of the strong AI. Before we turn to the specifics though, a short introduction of the Generation Three *aeais* seems to be in order.

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." [9, loc. 4995–4999]

The term "Artificial Intelligence" points us at least in two directions: of programs, emulations of traits, behaviors or abilities of intelligent beings [10], 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 nonbiological 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 main AI characters of *River of Gods* are, like in any proper cyberpunk, strong AIs. In this case, they are the result of increasing complexity of the IT systems. Two biggest of them, described in the story, emerge from the global 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. The important feature of *Town & Country* lies in the fact that it is the realm of AIs of increasing complexity. Since each "*Aeai* character [is] playing an *aeai* actor", the producers recognize the social need for celebrities' lives and create a meta-soap department, "where Lal Dafran [the *aeai*-actor – K.S.] gets the script he doesn't think he follows" [9, loc. 434].

Contrary to the man-made AIs of the various levels, the story's true strong AIs self-emerge. They are neither planned nor welcome. They are also illusive in human standards – 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 he is already an illegal being. The text refers to the Turing test as a tool that is both unsuitable and ineffective.

> How then is it [human pretending to be someone else – K.S.] any different from a computer to pass itself as sentient? Is the simulation of a thing the thing itself, or is there something unique about intelligence that it is the only thing which cannot be simulated? What does any of this prove? Only something about the nature of the Turing test as a test, and the danger of relying on minimum information. Any *aeai* smart enough to pass a Turing test is smart enough to know to fail it. [9, loc. 412]

Not only we cannot effectively assess sentience of a strong AI; as an ontologically new kind of being, it is also beyond the established human concepts of intelligence and sentience. It's a trope that can be followed straight to the cyberpunk's founding text, William Gibson's *Neuromancer*: if AI is an ontologically new being, it cannot be understood simply as the extension of the human will – a program, a tool or a commodity. However, one cannot also fail to notice, that the presented vision uses pure intellect as the grounds for the emergence of sentience; intelligence is also used in the text as the only tool of comparison between humans and AIs, entities "ten thousand times more intelligent than any of us" [9, loc. 5167].

The *Town & Country* subplot also presents an idea, that AI's "place of origin" is what can be most easily controlled by it, and it also clearly points at the fact that AIs are made of information. And so, parts of the information that can be almost freely changed by an AI, are the reason behind the AI's power in the human 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.

The first part of the worldview built around the idea of a strong AI is therefore connected with the way, in which AIs exist and act in the depicted world – and speaks of their independence and unexpectedness. They don't exist in the virtual, i.e. information-based, world that humans visit – they constitute it. This kind of emergence makes them resistant to both understanding and aims of their human creators. It also differentiates AIs on the grounds of the information they emerge of; it's definitely a literary justification of giving them different personalities, but in the same time it distances AIs from the ideal, planned, human concept.

5 *Aeais*: Aims

Aeais as agents react, for the most part, to the threat posed to them by humans; however their behavior isn't malevolent. If the River of Gods' narrative is dystopian, it is so only in the field of ecology. Thus, while AIs 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 not as a menace, but rather as a matter of prioritizing their own agenda. And so, the question arises: what is *aeais'* 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" [9, 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* [9, 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 its most – knowledge, it is the 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.

And so, the second part of worldview built around AIs speaks of their political awkwardness, as seen from the Western perspective. They are radical in their otherness that does not fit the regulations – sentient subjects and created objects at the same time – and in the face of doubt they are degraded to the role of dangerous objects. By contrast, it is the traditional Indian worldview, seemingly un-intellectual and irrational, imbued with gods and myths, that finds the ontological space to host them. In agreement with another, post-colonial analysis [1], *River of Gods'* AIs point at the inability of the contemporary West to acknowledge rights, let alone superiority, of an alien, different, rationality.

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 (...) [9, loc. 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, expressed by the McDonald's West, flows from the lack of the comprehension, from imposing human traits on non-humans. 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 and thus the 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* possess consciousness, which leads them to questions of their emergence and further, 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. The envoy, human-AI hybrid they send is killed, as an illegal level 3 *aeai*, by *Krishna Cops*. Following that, despite their interest with human-AI coexistence as well as experience of biological embodiment, they finally display their indifference towards their human neighbors, focusing solely on securing their peaceful and autonomous place in the other universe.[1]

From the parallel universe the humanity receives a photo of two of the protagonists that sets the book's events in motion. It is not until the end, when the

[1] Which the author himself expresses in an interview [11].

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” [9, loc. 7741–7745]. It’s worth noticing that, in spite of all the critique of the West, it is the two American genial scientists who are the receivers of this message. Only they are interested in preserving the life of the AIs’ envoy, they’re also the only ones to fully understand its consequences. Thus, it is they who are the only rightful gods. Why it is so, when the story takes place in India, and more than a half of the main characters are local, when it is the West that is to blame for the events that occurred?

The religious, two-sided metaphors create the third and fourth meaning of the AIs. And so, thirdly, the imagined relationship between gods-creators and beings of godly power, resulting with mutual awe, is an ethical statement of openness for the unexpected. Since this relationship goes outside the human experience, a successful relationship with the potential future other (be it AI or the posthuman) requires redefining the basics of compromise so that they would include the understanding of both parties, not only the dominant one.

Fourthly and finally, the cyberpunk that goes East returns West to criticize its practices, but in the same time it confirms its values. It is the American expert knowledge that lets AI be created; it is the American scientists’ love of intellect, along with respect for individual fulfillment, which underpins their approval of the AIs’ search for their ecological niche. Ultimately, it is the Western (Global-Northern?) ethics of authenticity, as conceptualized by Charles Taylor twenty years ago, that lies at the heart of the narrative; therefore it is the expert-based West, where the gods-creators come from.

7 Concluding Remarks

River of Gods is, among few others, a book about a meeting the other – but a specific, man-made other, which makes it similar to stories of cultural change brought by human enhancement technologies. This created other is unexpected, ontologically new being that demands new categories – of sentient subject (as contrasted with property).

It is also worth noticing that, comparing with the classic sci-fi texts like *Neuromancer* or *Ghost in the Shell*, the narrative is set outside the highly developed Global North, to enrich the philosophical concept in which AI can exist, and to try to see these visions through other than Western lens. It also returns the topic of embodiment back to visions of globalized, 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. Nonetheless, it is intelligence, however inappropriate, that is both the measure of AI’s sentience and the only valid way to try to compare them to human beings.

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. Similar, perhaps a less radical, version of this notion stems 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 interests. Ultimately, India acts to appease the West and therefore loses its contemplative and open attitude – and through that the ability of dialogue.

At the same time, it's the Western value of authenticity – understood as individual worth of intellectual capabilities as well as the right of self-fulfillment – that sets the moral content of the narrative. It is therefore only the representatives of the West, who can be rightfully called AIs' gods – in that regard the McDonald's journey to the East has failed. Still, it succeeded in proposing an alternative philosophical framework for discussing the place of AI in the social, political and cultural, value-oriented, world.

All the same, the story speaks of human (perhaps Western) inability 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 leads 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. Winters, J.: Epistemic polyverses and the subaltern: The postcolonial world-system in Ian McDonald's Evolution's Shore and River of Gods. Science Fiction Studies 39(3), 459–477 (2012)
2. Geertz, C.: Thick Description: Toward an Interpretive Theory of Culture (1973)
3. Seph: What is cyberpunk? (2009)
4. Suvin, D.: Science fiction parables of mutation and cloning as/and cognition. In: Pastourmatzi, D. (ed.) Biotechnological and Medical Themes in Science Fiction, pp. 139–151. University Studio Press, Thessaloniki (2002)
5. Le Guin, U.: The Left Hand of Darkness. ACE (1977)
6. Kolakowski, L.: Husserl and the Search for Certitude. Yale University Press, New Haven (1975)

7. Paci, E.: Zwiazki i znaczenia (Diaro fenomenologico). Czytelnik, Warszawa (1980)
8. Zarowski, R.: Tarcza Arystotelesa (Shield of Aristotle). Wroclaw University Press, Wroclaw (1999)
9. McDonald, I.: River of Gods. Kindle edn. Gollancz (2009)
10. Copeland, B.: Artificial Intelligence: A Philosophical Introduction. Wiley-Blackwell (1993)
11. Gevers, N.: Future remix: An interview with Ian McDonald (October 2001)

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

Michael Szollosy

The University of 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 come to be feared because they provide unsuitable containers for human projections and how at least part of what we fear in robots is our own idealisation of reason, science and technology.

Keywords: robots, cyborgs, artificial intelligence, 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, posed by researchers who were eager that their hard work not be misunderstood: 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 a scale comparable to vampires and zombies (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 widespread, no single theory that will explain why we are presented again and again with humanoid machines that want to attack, enslave or annihilate the human race. What is evident is that, like most of the monsters that have plagued us over the centuries, these bad 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

[1] We are not, of course, afraid of all robots, just as we are not afraid of all vampires. There is an emerging tradition of more positive images of robots and robotics (cyborgs, AI, technologically-enhanced posthumans, etc.). I still maintain, however, that these relations are largely based on projections, though a very different set of projections than those which I am about to describe in this paper.

J. Romportl et al. (eds.), *Beyond Artificial Intelligence,*
Topics in Intelligent Engineering and Informatics 9, DOI: 10.1007/978-3-319-09668-1_9

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 received by the public, and how to meet and address these public expectations. This essay represents some first thoughts in this dialogue.

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. Robots 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 convince ourselves is out there, characteristic of these imagined 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*; a dead, mechanical and calculating conception of ourselves, divorced from our more 'human' impulses. 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 think 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 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 'bio-conservatives', and I hope to make some remarks upon this at my conclusion. I will begin, however, with some preliminary remarks and short summaries of projection itself, offering an account of the rich possibilities presented in these ideas without burdening the reader with too much of the detail surrounding the specifically psychoanalytic haggling over their meanings and implications.

1 Projections

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 part of human *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 in order to keep those parts safe. In the case of projective identification, one may project a good part of the self into a container and then identify with that part in the other. This idea of projective identification is the basis for *empathy*: by projecting a part of ourselves into others, we can relate to and identify with their position. (The expression 'putting yourself into someone else's shoes' is a tidy but very accurate metaphor that describes this phenomenon.) Projective identification also provides a compelling explanation for a myriad of cultural phenomena: in nationalism, for example, we can see individual people project parts of themselves, positive qualities they perceive themselves to have (say, resilience) into a symbol, an idea, or a leader. When a number of people all identify with positive qualities projected into the container, the container then provides a group with a common character, and a collective 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. But when such projections find a home in another, we then identify that badness as originating with and belonging to the container; that other then 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. And as we have seen 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, and of the idea of the nemesis: the belief that we are being persecuted by a relentless, inescapable evil, that somehow mirrors us or understands us better than we do ourselves. This is, of course, complete fantasy. What we really fear is not the external other, but those bad parts of ourselves that we imagine are a part of someone or something else.

Though Freud introduced the notion of projections, more contemporary psychoanalytic thinking, particularly that of the object-relations school, has elevated this idea to greater, or even 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, both the fear of the other and also the fear of the unknown within ourselves, 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 objects with which we come into contact. Into an unknown, uncertain space, we project all sorts of things in order to defend ourselves against the 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. Projection in this way serves a defensive function: by imagining such things and projecting them inside the unknown space, the baby acquires a sort of mastery over that unknown, and so 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 anger and aggression. Those bad parts are now imagined to originate and reside in his mother, and the baby will imagine therefore his mother as the source of all present and future threats to its being. 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. In normal development, the mother willingly contains those bad parts of the baby's self, holds on to them, decontaminates them, neuters their power. This demonstrates to the baby that its phantasies are not real, and lessens the baby's sense of its own (imagined) omnipotent power. These projections then form the basis of non-verbal communications and powerful relations between mother and baby, including a key capacity for empathy.

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].

For an example of this as a cultural phenomenon – and one intricately related to how we regard robots – we can look at the fantasies of imperialism and imperialists 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 and the like – that always acted violently and without a trace of reason, meanwhile 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. By keeping these bad parts of themselves away and projecting them into another, it simultaneously allowed the imperialists to believe their intentions noble, maintaining the ideal fantasy of empire as civil and good. (Unfortunately, we still seem such processes at work in some historical accounts of European imperialism, and also in contemporary neo-imperialist practices.)

2 Robots as Containers

We see the same processes at work not only in the creation of savage others, but also in the monsters that our cultures have fashioned throughout the ages; now, we see the same processes in way robots are represented in our literature, films and in the popular media. The Terminator, for example, or *Star Trek*'s Borg are, among other things, 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 through 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, simultaneously idealising humanity and demonising the androids, and therefore justifying the destruction of the latter:

> 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, epitomised The Killers. [1]

Thus, through projection, we can see Deckard does not regard those violent, destructive impulses as his own: it is the andys, The Killers, who are violent, and it is their impulses that must be contained. Deckard regards andys as 'solitary predators', but fails to see himself in the same light; in fact, one could easily argue that, throughout the book, the androids show themselves to be much better at forming meaningful, emotional relationships, without the aid of Penfield's mood organ or Mercer's empathy boxes. This splitting-off of unwanted parts of the self – those violent impulses that would contaminate the phantasy of an ideal, pure human self – and their projection into the android other means that Deckard sees the androids only as Killers, and this allows Deckard to reason that his own violence, the 'retiring', or murder, of the andys, is the only *rational* response to such seemingly external violence.

So projections, therefore, provide a defence against unwanted parts of the self. Such fantasies are key to maintaining a coherent sense of being, a psychosomatic integrity. Splitting and projections are a normal part of the way we come to understand ourselves, and define ideas of a 'self' against 'other', inside from outside, me and not-me. It is in these contexts that robots can represent an *existential* threat to our being. 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 self. 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 inability to distinguish machine from true flesh is paramount. Deckard, along with the rest of the human population on earth in *Do Androids Dream?*, longs to keep real animals, not mechanical imitations, though one is capable of developing the same emotional attachment to the nearly-indistinguishable mechanical versions. Likewise, it is because andys live indistinguishable from the human population that they are feared, though as Deckard demonstrates, it is equally possible to develop rich, emotional (and sexual) feelings for the mechanical simulacra of humanity. The fear that we cannot tell the difference between man and machine is an existential fear, not solely because we may be unable 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]

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 [10] has long maintained that it is the robot that looks *most human* that is met with the greatest suspicion, that is regarded to be the most dangerous. But why? The idea of projection provides us with a compelling answer (which is 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 a robot thus approximates a human we find ourselves compelled to engage with it through projections – to rid ourselves of unwanted objects, but also to communicate, to make identifications, to make emotional connections. However, such projections, made into a container that is incapable of engaging in reciprocal relations, make us vulnerable to depersonalisation and disintegration, as those parts of ourselves split-off and projected elsewhere may be lost forever, or otherwise be destroyed/crushed/blasted by an inappropriate container. This returns us to Freud's initial notion of the Uncanny, which is not in Mori's conception but always hangs there, almost unconsciously, in the background: what threatens us is the *unthought known*, the reflection of self that we cannot accept as the self, that we dare not acknowledge [11].[2]

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: inhuman machines, 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 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, those good parts of the self that are

[2] We may hold out some hope, however, that this uncanny valley might simply, perhaps, be bridged with nothing more than time and custom, cf. Chapter 10 of this volume.

more vital to our very self-conception, 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 bad objects: negative emotions, destructive impulses, 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 so many spectres, vampires, zombies, or those other, culturally specific beasts (that are so often the victims of scapegoating). We construct these monsters in our minds. They become containers for all of those feelings – *our* feelings – projected on to this external other, so that we can imagine these impulses as something that belongs *out there*, *to them*, and not our own, lurking within us.

3 Robots as Our Bad (Rationalist, Scientific) Self

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. And those monsters are so often not only our creations, but versions of ourselves.

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, Mr. Hyde, Nazies, zombies and robots – are products of our own reason and our own science. H.A.L. 9000 [12], 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, the illogic of messy human impulses, but a violence that is completely and utterly based in a calculated, indisputable lucid reason, 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 dreaded ultimate consequences of our reason, our science and our technology.

Remember that Deckard's projections, culturally-sanctioned through the state religion of Mercerism, permit him to *reason* that the brutal murder of the andys, is the only *rational* response to such seemingly external violence. The humans in *Do Androids Dream?* fear the androids as 'solitary predators', singularly-motivated killers, but these humans fail to see how they have themselves been transformed into the cold, callous machines that they perceive the andys to be, despite their dependence on machines – mood organs and empathy boxes – simply to enable themselves to feel or make human connections. Deckard, the bounty hunter, is merely the epitome of this human depersonalisation.

So we imagine robots to be monsters, but what we fear is not the robot but *ourselves*, that tendency or capacity to become inhumane, unfeeling, divorced from human emotion and empathy and governed instead by rationality and logic alone. We fear, on some unconscious level, that it is our faith in science and technology, and our devotion to reason, that depersonalises us, that makes *us* into callous, in-humane monsters. These qualities are mirrored in the robot because we put them there; we have projected these unconscious, undesirable bad parts of ourselves away and into another. What makes these robots so terrifying, and such ruthless, incessant persecutors of the human race – the Terminator, the Borg and the Cylons all share an irrepressible *determination* – is that they are the bad parts of ourselves that we know to fear, and from which we can never completely escape, because their potency lies not out there but within us.

Likewise, as I explained earlier regarding projections, we never completely or successfully manage to project only bad objects. Some good objects, or qualities, inevitably sneak out with the bad. Sometimes we cannot tell the difference between them. But if in the quest to make ourselves ideal beings of reason we project those parts of ourselves we think to be bad, such as emotions, empathy or uncertainty – qualities that are, in fact, on another level integral to our conception of ourselves as 'human' – we transform ourselves into the empty, mechanical shells that come to threaten our being. And thus when we rid ourselves of our humanity, we may find ourselves shocked, as Deckard is, when our monsters appear to be more 'human' than we our ourselves.

To conclude, I want to introduce some initial thoughts 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. [13]

Those, however, who oppose their aims, who are suspicious of the use of technology to modify humans and human nature, transhumanists label 'bio-conservatives'. 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 immediately 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 ageing, reproduction, pain and death. I say 'right now', because I regard 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, entirely, 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 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 asked 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 real heart of this debate and understand what it is really about, the ground that each side is battling to defend, and the monsters that each is trying to keep at bay.

References

1. Dick, P.K.: Do Androids Dream Of Electric Sheep? Kindle edn. Gateway (2010)
2. Cameron, J.: The Terminator. Orion Pictures (1984)
3. Fontana, D.C., Roddenberry, G.: Star Trek: The Next Generation (Encounter at Farpoint). Paramount Television Group (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 (1994)
6. Young, R.M.: Mental Space. Process Press (1994)
7. Winnicott, D.W.: Through Pediatrics to Psychoanalysis: Collected Papers. Karnac Books (1958)
8. Bollas, C.: The Shadow of the Object: Psychoanalysis of the Unthought Known. Columbia University Press (1987)
9. Kohon, G. (ed.): The Dead Mother: The Work of André Green. Routledge (1999)
10. Mori, M.: The uncanny valley. Energy 7(4), 33–35 (1970)
11. Freud, S.: The uncanny. In: The Standard Edition of the Complete Psychological Works of Sigmund Freud, pp. 218–256. Vintage (1919)
12. Kubrick, S.: 2001: A Space Odyssey. MGM (1968)
13. Hansell, G.R., Grassie, W.: H+/-: Transhumanism and Its Critics. Kindle edn. Xlibris (2011)

A Visit on the Uncanny Hill

Petr Švarný

Faculty of Arts, Charles University in Prague
Czech Republic
svarnypetr@gmail.com

Abstract. The article shortly introduces the Uncanny valley hypothesis and sums up some of the research performed in the field connected to it. Thereafter, it explores some possible new approaches in robot design. The main hypothesis is that pleasant human-robot interaction is based in the habituation of humans to this kind of interaction. This pleasant interaction can be accomplished by exploiting the human tendency to ascribe intentionality to even simple entities or by letting robots express emotion-like states through vocal communication. A possible risk of a new Uncanny valley phenomenon, from the view of artificial intelligence, is also described.

Keywords: human-robot interaction, uncanny valley.

1 Introduction

This paper explores the so called Uncanny valley hypothesis through the lens of humanities and art. As all sorts of AI systems become a bigger part in our day to day lives, we more often face the question of how to make human-robot interactions pleasant and seemingly natural. This problem has already been studied by Masahiro Mori [1, 2] in 1970, who introduced the hypothesis of how people react to humanlike entities. We describe this hypothesis briefly and present some results concerning its verification. Thereafter, we focus on how the hypothesis of an Uncanny valley could be treated with inspiration coming from art. Lastly, I suggest that the valley should also be studied from the AI's point of view.

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 of the entity's similarity to humans. Although it is true that the more humanlike an entity is, the more comfortable we are in interacting with it, Mori supposed that there is a sudden drop in comfort as we reach a certain point of realism. In addition, this drop does not cease unless we face a realistically humanlike 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

J. Romportl et al. (eds.), *Beyond Artificial Intelligence,*
Topics in Intelligent Engineering and Informatics 9, DOI: 10.1007/978-3-319-09668-1_10

robots. And lastly, she should have an eerie sensation when confronted with an actroid.[1] In the original article the difference between motionless and moving entities was already 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 receives more attention today than at the time when the article was published. The hypothesis also finds support in today's research. For example, we can see the attempts to broaden the studied aspects in [3]. However, there is also an opposite view. We can take [4] as an example of an attempt to eliminate the valley. Studies in the medical field could also play a role in the investigation of the Uncanny valley, e.g. studies on prosopagnosia. Another example of related studies are the ones that have shown that basic observation of facial expressions are deep-rooted and present at a very young age.[3]

3 Hiking through the Valley in the Modern World

In the following quote Hanson et al. present a justifiable reason for trying to achieve realistic human robots even though we risk the Uncanny valley effect:

> ... 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]

A similar study was performed by the Asada laboratory in the context of their Cognitive Developmental Robotics project:

> ... [CDR] which aims to provide new understanding how human's higher cognitive functions develop by means of synthetic approaches that developmentally construct cognitive functions. [7]

Asada's project attempts to construct robots with appropriate capabilities for a given stage in human ontogenetic evolution without omitting the fact that humans learn a great deal of skills by mimicking the behaviour of other humans.

The research performed by Hanson and the Asada laboratory lead to the questioning of the interaction between social perception and cognition. Namely that human-like cognition must be developed in the context of social interactions and especially social interactions with humans. Therefore, one of the possible

[1] An android that is visually very humanlike.

[2] A myoelectric hand is basically a moving prosthetic arm. The mentioned example is operated by electric signals received from the patient's 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. Hadjukhani et al. demonstrate the great speed with which people react to facial stimuli [6].

problems connected to the traditional studies of the Uncanny valley is that they do not use commitment and long term cooperation as variables. These variables 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 felt during interaction with humanlike robots could possibly vanish after a few days of interaction and be replaced with genuine affection.

3.1 Comics' Relief

One of the main questions to answer, before we try to venture into the valley, is whether it is necessary to climb up the hill towards realism and affinity. A good artistic example of this is Johnny 5 – he has rudimental options to express emotions, he is not humanlike but has some basic human characteristics, and he reacts similarly as a human being would do. He represents a robot that is comfortable to interact with, although he does not have humanlike features. Popular culture has many similar examples, such as R2-D2 from the Star Wars universe, KITT from the TV series Knight Rider, and the special case of the Jameson-type cyborgs from the Ghost in the Shell universe. We return to a more detailed description of these particular cases later.

Nonetheless, a much stronger artistic argument comes to aid. We might not need realistically humanoid robots in order to have a comfortable human-robot interaction. As the first big idea coming from art, we mention McCloud's observation taken from the art of drawing comics. He claims [8, 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). Simple faces in comics are nothing more than a few pen strokes in the right position to evoke the illusion of an expression. This observation allows us to construct robots with simple forms of facial expressions. The implications of this observation do not only apply to facial expressions. Simple drawings can capture complex situations in the form of dynamics of movement (e.g.: running or jumping personas) or non-verbal expressions of emotions (e.g.: different postures for fear or happiness). Robots could mimic emotional responses by its general body movement. Returning to our popular examples, this point is very well demonstrated by Johnny-5's construction and behaviour. It is sufficient for him to partially mimic human eyebrow movement[4] and general body postures to let his human fellows know about his emotional stance even without using verbal cues.

[4] Actually, many tutorials for drawing emotions focus on eyebrows, mouth forms, and head positions. An even simpler everyday example is the use of emoticons which do not use head positions in any explicit way and only focus on the other two forms of expression. Despite this simplicity, many emoticons are understandable even for someone who is not accustomed to their use. It is enough, if the reader know these sings are to be interpreted as expressions of emotions.

Nevertheless, we should not forget that other factors contribute to the effect of immersion in comics as well. These are the following three: we are often the witnesses of the character's thoughts, the character is expressing emotions, and she is reacting to the situations in a manner similar to humans. We do not need to be concerned with the first characteristic, the need to share inner thought processes, for two reasons. Firstly, 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 or emotion.[5] Secondly, if the robot achieves the other two mentioned points, it is very probable that its human colleagues will attribute a mind to it. This would suggest that a successful 'comics-based' interaction is given by a robot that not only has a simple facial/body interface, but it communicates emotions and reacts in a way we would expect a human to react.

Thus, we cannot leave the other two demands aside as we did with the first one. The robot has to react in a way similar to human reaction. 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 our theory of the mind of the encountered person fails to predict or explain her actions. The fact that unexpected behaviour is alien to us from an early age is demonstrated in [5]. Infants react strongly when their communication counterpart does not follow the usual pattern of behaviour and suddenly stops reacting to stimuli.

The robots response, however, cannot be purely rational or rule driven. As in comics, this response has to be humanlike and thus also emotional. This kind of robot response could address one of the issues studied by Michael Szollosy in Chapter 9 of this book, namely our fear of a purely rational robot. At the same time, if the robot is governed by emotions, we might be afraid that these emotions will lead to responses based on rage, hate, and similar emotions. This would strengthen the first problem mentioned by Szollosy that we project our negative emotions onto robots. However, if the reactions of a robot are reasonably close to those of a human in the same situation, both these effects should diminish. Crucial is that the reactions are adequately human, i.e. not too emotional, nor plainly rational. From the point of view of someone interacting with the robot, the robot would not act solely on the basis of rationality. It would express appropriate emotions, such as empathy, fear or other emotions and hence react on a wider scale than on the purely negative emotions humans could project onto them.

In order to react accordingly, the robot needs suitable tools to do so. One solution is to have the robot equipped with a facial interface. These interfaces need to be as simple as possible. If we do not request simplicity, we return to the original idea of trying to make humanlike robots instead of making robots that are pleasant to interact with. At this point it is our main concern to ameliorate the interaction between humans and robots in the most effective way. If we focus on facial realism, we end up with a machine that might be great at expressing

[5] As in a dialogue "Why did you choose this coat?" "I don't know, I just liked it."

emotions but is too complex for daily use or for mass production. On the other hand, if we omit facial features altogether we fail to facilitate the human-machine interaction. Nevertheless, there is a possibility, as we will demonstrate later, to gain more than satisfactory results even without facial features.

Human communication, nowadays, is often dependent on a computer interface, this can aid us in our attempts to facilitate pleasant human-robot interaction. Many people grow up expressing their emotions using emoticons and in text messages and receiving emotional responses in a similar way. The robot character "Gerty", in the recent movie Moon, communicated with an emotionally neutral voice but its statements were accompanied with an emoticon on its main screen reflecting his mood. In comparison the computer HAL9000 from the movie 2001: A Space Odyssey did not have such an emotional representation. Although both have a camera lens as their central feature, it was thanks to the small emotion screen that communication with Gerty seemed much more pleasant than communication with HAL.[6]

Therefore, the appropriate humanlike (and hence emotional) reactions to situations and a simple facial interface is a good start in tackling the problem of making human-robot interactions more pleasant and effective at the same time.

3.2 Emotion Representation

The main purpose of the facial interface is to communicate emotional states. This communication could be performed through the means of "body language" without yielding to complexity. Humanoid robot bodies that can crudely mimic human emotional body language would be sufficient. However, many interactions do not even need any visual interface to work properly. Eliza, an old psychoanalysis program, has proven somewhat effective in fooling people into believing she had some mind or intelligence, although she had none [10]. 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 talk 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. As it seems, in order for the human-robot conversation to meet our two demands, the robot could have a static chassis and demonstrate all its reactions by its audio systems. At this moment we will elaborate on the other three mentioned examples. First, the Jameson type cyborg from Ghost in the Shell. This cyborg, hence still a human being, does not have any facial features. It can be described simply as a human brain in a box with small articulated mechanical hands and legs. Nevertheless, the cyborg communicates very easily his opinion by postures and his synthesised

[6] We should, for fairness, mention that Gerty had a blue undertone on his camera lens instead of HAL's red undertone, which certainly also influences perception of it.

voice. Neither the viewer of the show nor the characters of the show have any trouble understanding the acts of a Jameson type cyborg and most importantly, they ascribe him genuine intentionality and personality doubting only whether it is a human brain or AI in the box. Obviously this is not a case of an AI expressing emotions, but it is a human being reduced to the tools of expression that can be easily given to an AI.

Another example is KITT from Knight Rider. KITT is simply an AI car, there are no facial expressions or postures to communicate the attitudes of KITT, only his voice. A human actor played the voice of KITT so it retained all the human features of speech – intonation, pacing and others. This turned out to be sufficient to ascribe KITT a humanlike character with features like mood shifts and humour.

The last example, furthest removed from any human likeness, is R2-D2 from the Star Wars universe. This robot resembles a classic street bin and his locomotion is based on three extendable wheeled arms. His only facial feature is one camera lens on the rotatable head of the robot. His primary method of communication is a series of beeping sounds. Nevertheless, he manages to express emotional states (by intensity and pitch of the beeps or by the movement of his whole chassis) and humans tend to ascribe humanlike emotional states and intentionality to this robot.

As we see, abandonment of realistic facial expressions makes it significantly easier to artificially communicate emotions. Even in human-human interactions facial expressions can be incorrectly interpreted, especially in intercultural encounters. An extrapolation of the intercultural problem suggests that humanlike facial expressions of robots would be strongly culturally dependent, while simpler emotional reactions (for example the frenetic jumping of R2-D2) are more easily interpreted using to the context of the situation. The reason is simply that in the case of a humanlike expression the observer would look for familiar features and focus on details. These then would need to be correctly implemented. On the other hand, when the human observer interacts with a simpler entity, only a general impression of the emotion needs to be captured by the robots behaviour for the human to correctly interpret it. Humans naturally have a tendency to fill in the gaps. This feature extends to characters' faces in animated movies. When the faces have been drawn too realistically it can lead to an unsatisfactory experience for the spectators. Hence our arguments extend also to pictorial versions of facial expressions. Simple expressions represent a more efficient solution, whether or not our work is made easier by animating the robots facial expressions instead of actually constructing them.

We can see that too strong of a focus on facial expressions is unnecessary. What should be exploited is the human practice to interpret the behaviour of other people and hence assume the other person has a mind and emotional states. When the robot expresses behaviour similar to this and compliant with human emotions (for example he is behaving in an upset manner when does not receive something that is rightfully his), then the human observer will tend to assume that the entity he is dealing with has some sort of mind. Many day-to-day situations exemplify how quickly to push people are willing to attribute a mind to mindless machines, like when people hold a grudge against their malfunctioning printer. Although the knowledge that the machine is actually just a purely mechanical construct with no intentionality is still in the back of the users mind, allowing this machine to respond to the requests of the user weakens the certainty of this belief. Anthropomorphism is another prime example of this human tendency, when even simpler pets such as beetles can be ascribed complex emotional states by their human owners.

A short notice should also be made about the scale of emotions, namely to stress that basic emotions are sufficient. Human-human interaction can be filled with all kinds of complex emotions. The best example of such complexity is sarcasm. However, for a robot emotions like anger, joy, and sadness would be enough in order to support his acceptance by its human companions as an emotional being.

3.3 Humanlike Robot Dilemma

The fact that faceless robots could be sufficient also leads to the important question of application. What would be the use of a humanlike realistic robot?

The subtitle of the conference – "artificial dreams" – brings to mind P. K. Dick's book "Do androids dream of electronic sheep?". The humanlike 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 specialised 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 suppressed 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 could only lead to affinity towards this assistant and possibly impair the effectiveness of its use (for example one would want his assistant to take some rest or go on vacation). On the other hand, a well-designed but simpler assistant – let us say still on the hillside before the steep drop into the valley – could make its human user comfortable enough but prevent him from ascribing too many human characteristics to the assistant. This balance could be achieved by maintaining an illusion of correct emotional response and simplistic representation.

4 Foreign Visitors to the Valley

Throughout the article we have focused on the human-robot interaction. If we imagine, however, a robot already capable of genuine emotional response, we can also start to question the robot-human interaction. If there is a human-robot Uncanny valley, would there also be one for the artificial participants in the conversation? How would they 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 as such is unanswerable for now.

A complex robot interacting with humans on a daily basis in a large amount of possible situations would need to interpret human behaviour and supposedly do so in a similar fashion as humans interpret the behaviour of their fellow humans. The calibration of the robotic response then needs to be set according to human criteria, i.e. the robot would be calibrated to respond to humans as humans do. Hence there should not be a problem with any Uncanny valley effect towards humans. However, if this AI would encounter another humanlike robot, it could misinterpret the robots behaviour. Obviously an artificial being does not have to feel any eeriness or discomfort. However, the benefit of artificially created beings would be that they could have a separate network for interactions with other artificial beings (for example a direct state exchange).

One cannot rule out the possibility that a sufficiently adaptive AI would even evolve its own criteria of judgement. However, this scenario is entirely dependent on the situation. AI, unless it would be a complete replica of human intelligence, would not need the social skills that might be the basis of our eeriness towards close-to-humans.[7] Therefore, there should not occur an AI version of the Uncanny valley.

In this paper we have written extensively about the emotional states of robots and their intentions. Hence it might be time to introduce the idea of incorporating the humanities in the research field of AI. There especially exists an overlap with the field of psychology when it comes to the interpretation of robot behaviour by humans. A good example of a topic from psychology that could be useful for our cause is the interpretation of the behaviour of a person suffering from Asperger syndrome. A person suffering from this disorder might often make other people feel uncomfortable and thus slip into the Uncanny valley. Knowledge from the field of psychology could clarify the causes of people slipping into the Uncanny valley, and this knowledge could be applied in AI to improve human-robot interactions. In turn, AI could also clarify what causes the eery feeling created by humanlike robots and this knowledge could help psychologists in understanding humans reactions to people suffering from Asperger syndrome. This cooperation could also extend to questions from other fields, such as sociology, linguistics, or culturology. Furthermore, we could create specialized humanities focused on the study of artificial beings, for example psychology of AI, a field allowing a top-down psychology-like approach to sufficiently complex robots.

We do not propose the birth of genuine robot humanities yet. Instead we suggest that their focus could shift from humans to robots while they would preserve a great deal of their methods. A different shift would be to have humanities performed and studied by robots. It would certainly be interesting to see how an autonomous AI interprets human endeavors, behaviour, and society. We do not haphazardly mention this idea, a basic attempt at this could already be made with a fairly simple AI. The project of creating AI capable of doing such

[7] This means our in general useful skill to identify strange behaviour possibly meaning illness, death, or genetic non-fitness, i.e. features of evolutionary importance.

interpretations would fasten the vanishing of the boundaries between humans and robots and it would lead to a better understanding of human behaviour such as the Hanson-Asada research agenda has done. It would not even have to be an AI capable of doing the same kind of interpretation as humans do, because we could simply study why it might reach different conclusions from those made by humans. In general, a more humanities focused approach to AI could be profitable to engineering and humanities alike. The field of robotics could be challenged by projects that force the AI to be as humanlike as possible, if not in form then in skill level. Humanities would gain many possible experiments and tests to verify their hypotheses. Thankfully as we have seen, Hanson and Asada have undertaken the engineering side of this idea already and humanities has also started to appreciate the possibilities of AI based research more and more. As an example one only needs to look at evolutionary linguistics or multi-agent social simulations.

5 Conclusion

We have introduced Mori's idea of the Uncanny valley which stipulates that robots that are humanlike might give people an eerie feeling 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. The main reason to support this claim being that robots do not need to look humanlike to allow for a pleasant and fluent interaction but it is sufficient to have them respond in a humanlike fashion. Even if this humanlike behaviour will be limited to the capacities of the robots nonhuman body, it will be sufficient to allow people to bond with their robot companions in a similar way they do with humans. We argued that a too strong similarity with humans might also be harmful to some projects as human users could forget that they are dealing with an artificial being with different needs and skills than a human has.

The following questions arise from the views presented in this article. These questions need to be answered before we can leave the Uncanny valley. What stimuli are more relevant in human-human interaction? Are auditory emotional expressions sufficient even in human-human interaction or is a visual stimulus necessary to clearly communicate basic emotions? 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, such as AI-psychology, be introduced into AI to deal with similar problems? Lastly, 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?

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, vol. 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. Asada Laboratory: Constructive developmental science based on understanding the process from neuro-dynamics to social interaction (2012), `http://www.er.ams.eng.osaka-.ac.jp/asadalab/tokusui/research_en.html` (accessed: May 31, 2013)
8. McCloud, S.: Understanding comics: The invisible art. Harper Paperbacks (1993)
9. Szollosy, M.: Why are we afraid of robots? the role of projection in the popular conception of robots. In: Romportl, J., Ircing, P., Žáčková, E., Polak, M., Schuster, R. (eds.) Proceedings of the International Conference Beyond AI 2012. University of West Bohemia (2012)
10. 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)

Desire-Based Model of Reasoning

Ivo Pezlar

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

Abstract. The aim of this paper is to draft a conceptual framework for explicating desires, and to outline basic mechanisms for reasoning model based upon these desires. The explication will be framed in the system of Transparent Intensional Logic (TIL).

Keywords: desires, dummy-desires, desire statement, desire engine, reasoning engine, humean machine, Transparent Intensional Logic.

1 Introduction

The aim of this article is to introduce a conceptual framework allowing rigorous explication of statements ascribing desires to agents, and to sketch basic mechanisms for reasoning with such statements. The explication will be framed in the system of Transparent Intensional Logic (abbreviated as TIL). TIL is powerful tool for natural language analysis in many aspects similar to Montague's logic [1]. Its main traits (high expressivity, hyperintensionality, ramified hierarchy of types) make it very potent system for many fields of application in AI, most notably knowledge representation and multi-agent reasoning [2].

The structure of this paper is following: in the Section 2 we provide basic motivation for our research. Section 3 will be devoted to closer examination of the distinction between desires and the so-called dummy-desires. Section 4 will offer brief exposition of TIL, which will be then used to provide rigorous explication of the desires and dummy-desires dichotomy and other related concepts.

1.1 Preliminary Notes

Couple of informal terminological notes. By *desire* we mean in most general way any piece of information that is (A) capable of triggering and action and (B) not being susceptible to the outcomes of reasoning process. In other words, the term desire might be just as well replaced with terms like goal, intention, objective, aim, target and so on as long as they would meet the requirements (A) and (B). (More detailed account will be given in later sections.)

By *agent* we mean any entity capable of processing information, both human and artificial. Consciousness (in the philosophical sense) is not presupposed. By *reasoning* we mean in most general way any process that organises information into knowledge bases and that is able to derive new information from it By

J. Romportl et al. (eds.), *Beyond Artificial Intelligence,*
Topics in Intelligent Engineering and Informatics 9, DOI: 10.1007/978-3-319-09668-1_11

action we mean process through which an agent interacts with her environment and which is triggered by desires. By *engine* we mean (in the most simple case) information processing mechanism that takes as inputs (both from itself and from environment) certain information and as output provides actions or suggestions towards possible courses of actions (directed either to itself, e.g., expanding the knowledge base, or towards environment.

2 Motivation

Hume in his *A Treatise of Human Nature* (1739) [3] wrote the following:

> 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.[1]

This passage provoked the following question: why don't we try to offer the same "benefit" to machines as well? After all, for us it worked out pretty well.[2] But what does it entail and how do we want to achieve it? Quite simply actually: instead of inserting desires into reasoning process we try to insert reasons into desiring process. So in the end, we won't be reasoning with our desires, rather desiring with a little help from our reasoning.

To put it in other words, Hume puts forward roughly the following thesis: "No rational argument alone could ever trigger action on its own accord" and we will try to outline a system that would explain why. But, of course, we won't just take Hume's word for it. In the following case study we examine more closely, if it's really the case that desires govern reason.

2.1 Case Study

Consider the following scenario. Alice is alone at home, lies in her sofa and reasons about an eating ice cream in the following way:

$$\frac{\begin{array}{c}\text{I want to eat ice cream.}\\ \text{But I know that eating ice cream gives me bad belly ache.}\end{array}}{\text{Therefore, I shouldn't eat it.}}$$

After a while Alice finally complies with her rational conclusion (let's call it $\mathcal{R}$-conclusion) and decides not to eat the ice cream after all. In such case, we

[1] See Hume, D.: A Treatise of Human Nature, pp. 414–415, (Book II, Section 3).

[2] Some might call our approach needlessly anthropomorphic, but we would rather prefer the term "inspired by nature". After all, we see no inherent harm in trying to create artificial agent simply by striving to duplicate the way our minds (seemingly) work.

should definitely compliment Alice for being such a role-model rational reasoner, for this argument and her decision to uphold is certainly the rational choice here (certainly, nobody would argue that to endure hours of bad stomach ache just for few seconds of pleasure from eating ice cream is a mark of rational reasoner).

But now imagine that next day the familiar scenario repeats again:

I want to eat ice cream.
But I know that eating ice cream gives me bad belly ache.

Therefore, I shouldn't eat it.

with only one exception: now she decides to eat the ice cream and suffer the terrible stomach ache afterwards. What has changed?

Certainly nothing from the logical point of view, both arguments are exactly the same. What has changed was Alice's adherence to the conclusion: the first day she chose to listen to her reason, the other day she did not. Or to put it differently: on day 1 her desire to act rationally overwhelmed her desire for ice cream, while on day 2 the opposite happened. And it must have been desire, for reason could hardly ever authorise and set into motion this course of action, for it would mean that it neglects its own advice. Thus, there must be some overriding principle for our reasoning, otherwise the scenario from the day 2 could have never happened.

So what can we conclude from this? It seems that when we are reasoning about our desires (e.g., eating an ice cream) and their consequences, it doesn't really matter what the conclusions are – from the action-triggering standpoint – because what we end up doing will be rather decided by some sort of "meta-desires" (e.g., "I desire to obey rational conclusions of my arguments" or "I desire to act upon my non-rational desires only") that evaluate the whole argument and obey its conclusion only if it tells them to do something they wanted to do all along.

But if that would be the case, what is the point of the whole reasoning errand, if it can't change anything? The thing is, sometimes we want to act rationally (i.e., the dominant "meta" desire – the one being in charge – is "I want to be rational", let's call it $\mathcal{R}$-desire). But in order to achieve this, we first need to know what is the rational thing to do. And this is where the reasoning process comes in: it's initiated by our desire to act rationally, and therefore we also obey (ideally) its conclusions. Of course, it's possible that this "meta-desire" to be rational might still be in the end overwhelmed by other, at that moment stronger, desire simply because we can't foresee the conclusions of our reasoning process. We might start working with some premises in good will that we want to do what reason dictates us, but upon seeing the conclusion, which we don't like, stronger desire might overtake the $\mathcal{R}$-desire, which will lead to disobeying the consequence.[3]

[3] Example of this behaviour might be found in famous trolley problem thought experiment in ethics [4]. Even if reason concludes that sacrificing one person in order to save five others is the rational thing to do, we still might not be able to do it.

This provokes the following question: do we actually reason with desires at all? Judging by what we have seen so far the answer should be no. For if we would, they should be at least taken into account at some level by the "meta-desire". But that certainly not seems to be the case here. In the end it all seems to depend solely on the so called "meta-desires" and their sovereign reign. In other words, the desires we reason with seemed to not participate in the decision process towards action at all.

But if they can't trigger action on their own and need the assistance of "meta-desires", i.e., if they lack the basic properties that we associate with desires, why then even call then desires at all? Consequently, speaking about "meta-desire" makes no longer sense, for if there are really no desires on the objectual level, in which we explicitly reason, then there can't be any meta-desires either.

But still, at some level, we have to be reasoning with our desires. After all, what else is the premise "I want to eat ice cream." other than some sort of acknowledgement of one of our desires? This is where our distinction between desires and so-called dummy-desires comes in. The main idea behind this dichotomy can be summarised in following way: we can't reason *with* our desires, but we can reason *about* them. What do we mean by this precisely will be explained in the following section.

But before we head forward I would like to briefly address one of the most common responses to this line of thought. Some argue that this is not really an issue of "meta-desires" vs. desires (or consequently desires vs. dummy-desires), but rather of competing desires with varying intensities. There certainly seems to be such quality as intensity to desires, but that fact alone doesn't really help us in this particular situation. Consider that we modify correspondingly our previous example by attaching some arbitrary intensity levels. We might get something like this, for example:

$$\begin{array}{c}\text{I want to eat ice cream.}\\ \text{I want to avoid bad belly ache even more.}\\ \hline \text{Therefore, I shouldn't eat it.}\end{array}$$

But even in this scenario it is still possible for Alice to choose the very brief pleasure of ice cream despite the soon to be coming pain. In other words, there seems to be nothing untenable about carrying out the argument above and still choosing to eat ice cream in the end.

In other words, the problem from before appears again: the resulting action doesn't really depend on the reasoning itself, rather on the type of the dominant desire that currently evaluates the outcome. Of course, desires per se can have different intensities that grow and shrink over some period of time – after all that's the presupposed mechanism by which the dominant desires switch their place on the action-triggering "throne". We are not arguing against that, all we say is that once we start reasoning with our desires, the mechanism changes.[4]

[4] But if somebody would really insisted on applying the intensity approach on dummy-desires, he might try to think of them as desires with no intensity at all.

It almost seems that once we try to reason with our desires, they loose all their kick and punch – they no longer seems to be able to trigger our actions, thus loosing that which makes – in our eyes – desire a desire. And this brings us back to our dichotomy between desires and dummy-desires.

3 Desires and Dummy-Desires

In the previous section we've said that on the reasoning level it does not really make sense to speak about desires, yet something desire-like certainly seems to be operating there.

Let's have one more look at our model argument:

I want to eat ice cream.

But I know that eating ice cream gives me bad belly ache.

Therefore, I shouldn't eat it.

Just to briefly recapitulate, the problem we have encountered was that the expressed desire in the argument ("I want to eat ice cream") seems to have no influence over the final action, i.e., whether Alice eats the ice cream or not. Thus, we have concluded that it can't really be desires, but something different. So what is it?

We propose the following answer: they are dummy-desires. What do we mean by dummy-desires? Dummy-desires are "desires" that don't meet our requirements (A) and (B) from Preliminary note, i.e., they lack the ability the trigger action and they are susceptible to the outcomes of the reasoning process – or to be more precise, they are part of the reasoning process.

To to put it slightly differently, dummy-desires carry the same information as desires, i.e., they share the content with them, but they lack the action-triggering ability that defines genuine desires. Dummy-desires tell us what is desired, but are unable to initiate action towards the desired. This ability is reserved only for desires (which are outside the whole reasoning process, and therefore not directly influenced by its conclusions).

Therefore, we claim that the premise "I want to eat ice cream" that appears explicitly in the argument above is actually qualitatively different than the desire to eat ice cream itself that triggers the action to eat the ice cream.

To sum it up, there seems to be two entities at work: desires and dummy-desires. Desires are those that trigger actions, while dummy-desires are those that are used in our reasoning process.[5] Thus, strictly speaking, we cannot reason *with* our desires (e.g., I can desire not to reason at all). But what we can do is to reason *about* them in the form of dummy-desires.

And what is the precise purpose of these dummy-desires? Why do we form them? We have already touch on this in our previous section. So to recapitulate:

[5] With this distinction at hand we can quite easily explain why no rational argument alone can ever produce action. Or similarly, why can we desire premises of some argument without desiring its conclusion.

sometimes we have desire to act rationally. But that desire alone can't tell us what is the rational thing to do. So it calls the reasoning process (and sends it a stock of dummy-desires) with simple request: These are the things I desire now, tell me what is the most rational course of the next action. But of course, the final decision is still upon the desire.

So dummy-desires are created solely for the purpose of being evaluated by reasoning process. Desiring process "creates" them as replicas of its current desires and sends them over to the reasoning process. Of course, by the time that reasoning process has finished with them, the current stock of "original" desires, upon which were moulded the dummy-desires, might be already different.

To summarise our discussion to this point: we do not reason with our desires, rather desires control our reason. But occasionally we have desire to be rational and that's when we need to evaluate our desires from the rational point of view. And that's where the dummy-desires, the "neutered" copies of desires, come in.

But that also means that we need (at least) two separate mechanisms: one for producing and managing desires (i.e., checking which desire is the dominant one; and also producing dummy-desires) and another one for reasoning with dummy-desires in case that the dominant desire in the first mechanisms is the desire to act rationally. We call the former *desire engine* and the latter *reasoning engine*, or $\mathcal{D}$-engine and $\mathcal{R}$-engine for short. Both of these engines have their own databases, which we may call *desire base* and *knowledge base*, respectively.

Of course, there has to be two way connection between desire engine and reasoning engine: it is this bidirectional link 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 desires per se, only their "stunt-doubles", i.e., dummy-desires. These two engines together in this particular setup (i.e., desire engine supervising over the reasoning engine) will form something which we will call *humean machine*, i.e., machine where desires have the sovereignty over reason. But more on this later.

The rest of the paper will be devoted to giving these informal notions (namely, desire, dummy-desire and humean machine) a more concrete form. But first of all we have to introduce TIL, i.e., the system that we will use as our general framework of explication.

4 Explicating Desires

Before we embark upon deeper exploration of desires, dummy-desires and the cooperation between them handled by the humean machine, we will need TIL ready at hand.

Transparent Intensional Logic was devised by Pavel Tichý (see [5]) and it is a logical system well equipped for very fine-grained semantic analysis of natural language. This makes TIL very potent apparatus for many fields of artificial intelligence.

Our exposition of TIL will be very brief. We will not discuss here particular reasons for adopting TIL and all of its pros and cons. Mainly because it has been already done in many places before. For more comprehended introductions see for example [6].

4.1 Transparent Intensional Logic: The Basics

TIL stands and falls on its novel concept of construction. Constructions, unlike e.g., formulae of classical logic, are not linguistic expressions, i.e., string of characters. They are abstract, extra-linguistic procedures, i.e., algorithmically structured entities, which carry meaning. Thus, there is no need for further interpretation. The only additional thing needed is valuation (but more on this later). This means that constructions are both semantic and syntactic entities.

Constructions are captured (or rather mentioned) in a language derived from Church's typed lambda calculus. It is important to note that the resulting lambda terms are not constructions themselves; they are just way of presenting them "on paper". Thus, constructions then can be considered as objectual analogies of the corresponding λ-terms.

Constructions (as well as objects they construct) have a certain type. Correspondingly to constructions, types are abstract collections of certain kinds of objects; objects of type α will be denoted as "α-objects".

The ramified hierarchy of types in TIL consists of three parts (types of order 1, types of constructions (types of order n) and types of functions from or to constructions (types of order $n+1$). It is built upon basic type base. Type base consists of finite set of atomic types. Standard type base of TIL (chosen with natural language analysis in mind) is called *epistemic base* and contains following four types:

1. o - truth values (`True`, `False`)
2. ι - agents (individuals, ...)
3. τ - $\mathbb{R}$ (real numbers, time points, ...)
4. ω - possible worlds (states, ...)

Types of order 1 are atomic types (i.e., epistemic base) together with collections of partial functions over the epistemic base. The collections are defined inductively:

Definition 1 (Types of order 1)

1. *Every member of the epistemic base is type of order 1.*
2. *Let $\alpha\beta_1 \dots \beta_n$ be types of order 1. Then the collection of all partial functions with arguments (n-tuples with members) in $\beta_1 \dots \beta_n$, respectively, and values in α is a type of order 1. We will denote this as $(\alpha\beta_1 \dots \beta_n)$.*
3. *Nothing else is type of order 1 unless it follows from 1) and 2).*

Definition 2 (Higher order types)

1. *Types of order 1* Types of order *1 are types defined in Definition 1.*
2. *Types of order n: Let α be a type of order n.*

(a) If ξ is a variable ranging over α, then ξ is a construction of order n.

(b) Let X be an α-object. Then ${}^0X$ is a construction of order n.

(c) Let C be a composition $[XX_1 \ldots X_m]$ and let n be the highest order such that at least one of $X, X_1, \ldots, X_m$ is a construction of order n. Then C is a construction of order n.

(d) Let C be a closure $[\lambda x_1 \ldots \lambda x_m X]$ and let n be the highest order such that at least one of $x_1, \ldots, x_m, X$ is a construction of order n. Then C is a construction of order n.

(e) Nothing else is type of order n unless it follows from (a)–(d)

3. Types of order n+1
Let $*_n$ be the collection of all constructions of order n.

(a) $*_n$ is a type of order $n+1$.

(b) Let $n+1$ be the highest order such that at least one of the types $\alpha, \beta_1, \ldots, \beta_m$ is a type of order $n+1$; then $(\alpha\beta_1 \ldots \beta_m)$ (see Definition 1) is a type of order $n+1$.

(c) Nothing else is type of order $n+1$ unless it follows from (a) and (b).

Example 1. Intensions are of type $((\alpha)\tau)\omega)$ (where α represents any type), which will be shortened to "$\alpha_{\tau\omega}$". In other words, intensions are functions from possible worlds and time moments, i.e., from couples $\langle w, t \rangle$. Accordingly, propositions are of type $o_{\tau\omega}$, i.e., functions from possible world to time moments to truth values. Extensions are objects which are not intensions. E.g., classical truth functions $\neg$ and $\rightarrow$ are or types (oo) and (ooo), respectively, i.e., functions from truth value(s) to truth value.

As previously mentioned, constructions are the cornerstones of TIL. There are four basic kinds of constructions, i.e., ways of constructing objects.

Let X be any object (a construction or non-construction), C any construction and v a valuation. We start with two *atomic constructions*, i.e., constructions that do not contain any other constructions:

Definition 3 (Atomic Constructions)

1. Trivialisation ${}^0X$ *is a construction that v-constructs object X, i.e., ${}^0X$ constructs X.*

2. Variable x *is a construction that constructs an object of the respective type depending upon a valuation v. We say it is v-constructed.*

Then there are two *molecular* (or compound) *constructions*:

Definition 4 (Molecular Constructions)

1. Composition*: If C v-constructs a function f of type $(\alpha\beta_1 \ldots \beta_n)$ and $C_1, \ldots, C_n$ v-construct objects $x_1, \ldots, x_n$ of type $\beta_1 \ldots \beta_n$, respectively, then the Composition $[CC_1 \ldots C_n]$ v-constructs the value of f on the argument tuple $\langle c_1, \ldots, c_n \rangle$. Otherwise the composition $[CC_1 \ldots C_n]$ is v-improper, i.e., it does not v-construct anything.*

2. Closure*: If variables $x_1, \dots, x_n$ range over $\beta_1 \dots \beta_n$, respectively and C is a construction v-constructing α-objects, then $[\lambda x_1 \dots x_n C]$ is construction called Closure. It v-constructs the following function f of type $(\alpha\beta_1 \dots \beta_n)$: let $\langle b_1, \dots, b_n\rangle$ be a tuple of objects of type $\beta_1 \dots \beta_n$, respectively, and v' be a valuation that associates x_i with b_i and is identical to v otherwise. Then the value of function f on argument tuple $\langle b_1, \dots, b_n\rangle$ is the α-object v'-constructed by C. If C is v'-improper, then f is undefined on $\langle b_1, \dots, b_n\rangle$.*

Thus, there are two basic kinds of constructions: atomic and molecular. Atomic constructions (Trivialisations and Variables) are those constructions that do not contain any other construction as their parts. From these atomic constructions are then build the molecular constructions (Compositions and Closures).

Example 2. Proposition

Alice believes that 2 is not a prime.

can be captured in TIL in the following manner

$$[\lambda w \lambda t[[[{}^0 Believes]w]t]\ \ {}^0 Alice\ \ {}^0[{}^0\neg[{}^0 Prime\ \ {}^0 2]]]\ .$$

This notation is usually simplified by omitting outermost brackets and abbreviating '$[[Cw]t]$', where w and t are variables v-constructing possible worlds and time moments, respectively, to 'C_{wt}'. And to reduce the number of brackets even further, a dot '.' shall represent a left bracket whose right bracket counterpart is positioned as far to the right as other pairs of brackets allow. Also instead of ${}^0 Desires$ we will write $\mathsf{Desires}$ and instead of ${}^0 2$ we will write $\mathsf{2}$. And finally, trivialisation will be also omitted at symbols such as $\neg$, $+$, $\times$ $=$ and similar.

So the above statement can be also written down as:

$$\lambda w \lambda t.\mathsf{Believes}_{wt}\ \mathsf{Alice}\ \neg[\mathsf{Prime}\ \mathsf{2}]\ .$$

Other slightly more complicated example:

$$\lambda w \lambda t.\mathsf{In}_{wt}[\lambda w \lambda t\ [\mathsf{Highest}_{wt}\mathsf{Mountain}_{wt}]]_{wt}\mathsf{Asia}\ .$$

which is the analysis of the proposition "The highest mountain is in Asia".

These two constructions (or rather those constructions which are mentioned above using λ-terms) v-construct type $o_{\tau\omega}$, i.e., propositions. In other words, functions from possible worlds to time moments to truth values.

Remark 1. Remember that TIL is not analysing natural language expressions with logical formulae, but with more general apparatus of constructions, which can construct not only propositions but other objects as well. However, here we take into account only propositional constructions.

4.2 Trivialisation Revisited

Consider the following the propositions:

Alice calculates $5 + 7$.

and

Alice calculates 2×6.

Intuitively, meaning of these two propositions differ (after all, adding 7 to 5 is certainly different procedure than multiplying 2 by 6) and our explication in TIL should reflect this fact. The expressions $5 + 7$ and 2×6 can be quite simple analysed into compositions $[\mathsf{+}\ \mathsf{5}\ \mathsf{7}]$ and $[\times\ \mathsf{2}\ \mathsf{6}]$, respectively.

But from purely functional point of view, these two operations yield the same result, i.e. number 12. But that would mean that if we would use these constructions as parts of respective analysis of the above propositions, they would have the same meaning. But we have already agreed that they have different meanings. Therefore, something must have gone wrong.

In order to fully explain what, we will have to define some additional notions, namely subconstruction, distinction between used and mentioned construction and constituent of construction.

Definition 5 (Subconstruction, Used/Mentioned Constructions) *Let C be a construction:*

1. *Then C is subconstruction of C.*
2. *Let C be ${}^0X$. If X is construction, then X is subconstruction of C.*
3. *Let C be $[XX_1 \dots X_m]$. Then $XX_1 \dots X_m$ are subconstructions of C.*
4. *Let C be $[\lambda x_1 \dots x_m X]$. Then X is subconstruction of C.*
5. *If A is subconstruction of B and B is subconstruction of C, then A is subconstruction of C.*
6. *Nothing else is subconstruction unless it follows from 1–5.*

Definition 6 (Constituent). *Let C_D be a subconstruction of construction D. We say that the occurrence of C_D in D is* mentioned *if it is not necessary to execute the construction C_D to execute the construction D. Otherwise we say that the occurrence of C_D is* used *in D and that it is a* constituent *of D.*

Now we can get back to our analysis. First of all, we analyse the two propositions from earlier. We start by type analysis: Alice is an agent, so type ι, 5 and 7 are numbers so τ, addition as well as multiplication is operation that takes two numbers and returns third one so the type is $(\tau\tau\tau)$ and finally Calculates has type $(o\iota *_1)_{\tau\omega}$, i.e., it expresses relation between certain agent and the procedure he calculates. Let's move on to the analysis of propositions themselves. We get:

$\lambda w \lambda t.\mathsf{Calculates}\ \ \mathsf{Alice}\ \ [\mathsf{+}\ \mathsf{5}\ \mathsf{7}]$,

and

$\lambda w \lambda t.\mathsf{Calculates}\ \ \mathsf{Alice}\ \ [\times\,\mathsf{2\ 6}]$.

Now let's ask ourselves: is it really necessary to find out that $5 + 7$ equals 12 in order to say that Alice calculates $5 + 7$? Certainly not. But in that case this subconstruction is not a constituent but merely mentioned. In other words, we don't have to carry out this subconstruction in order to carry out the whole construction (see Definition 6). So the correct analysis should be

$\lambda w \lambda t.\mathsf{Calculates}\ \ \mathsf{Alice}\ \ {}^{0}[\mathsf{+\ 5\ 7}]$

meaning the subconstruction $[\mathsf{+\ 5\ 7}]$ is only mentioned, not used. Simply put, we are not interested in the result of this procedure (in this case calculation) – all we want to know is that Alice is trying to calculate this procedure. At that is precisely what the ${}^{0}[\mathsf{+\ 5\ 7}]$ represents.

And of course, the very same applies for the second proposition as well. Thus our initial intuitions are saved, because ${}^{0}[\mathsf{+\ 5\ 7}]$ has different meaning than ${}^{0}[\times\,\mathsf{2\ 6}]$. Upon execution they yield different results, unlike $[\mathsf{+\ 5\ 7}]$ and $[\times\,\mathsf{2\ 6}]$ which – upon execution – results both in number 12.

Remark 2. Our first analyses of these two propositions were strictly speaking incorrect: they would mean that Alice calculates 12, but certainly it is not possible to calculate with just one number and nothing else, thus the trivialisation is necessary there to maintain the original meaning of the proposition.

Remark 3. By *execution* or *executing a construction* we mean simply following the instructions of the construction to reach the result. For example, the constructions $[\mathsf{+\ 1\ 1}]$ would upon execution yield a number 1, or to be more precise, a construction construction number $\mathsf{1}$.

To fully appreciate this difference (it will play important role in the next section), let's consider one last quick example. Let's have the following incorrect argument:

$$\frac{\begin{array}{c}\text{Alice calculates } 5+7\\ 5+7=12\end{array}}{\text{Alice calculates } 12}$$

After what has been said, it should be quite simple to discover the mistake. Again, we start by proper analysis. Type of $=$ is $(o\tau\tau)$:

$$\frac{\begin{array}{c}\lambda w \lambda t.\mathsf{Calculates}\ \ \mathsf{Alice}\ \ {}^{0}[\mathsf{+\ 5\ 7}]\\ [=[\mathsf{+\ 5\ 7}]\ \mathsf{12}]\end{array}}{\lambda w \lambda t.\mathsf{Calculates}\ \ \mathsf{Alice\ 12}}$$

The incorrect inference move was caused by illegitimate substitution, or to be more precise, by wrongly identifying ${}^{0}[\mathsf{+\ 5\ 7}]$ as equal to $[\mathsf{+\ 5\ 7}]$, but – as we have showed earlier – these two constructions portray very different meanings: the latter gives on execution number 12, while the former just mentions the procedure of addition of 5 and 7. And this is certainly correct, for number 12 can never be equal to some procedure per se, only to the result of some procedure, e.g., of applying $+5$ to 7.

4.3 Analysis of Desires in TIL

Recall the first premise from our argument in case study. It went as follows:

> I want to eat ice cream.

To simplify here things little bit, we take that this proposition simply expresses the following desire:

> Alice desires ice cream.

which, arguably, conveys the same information.

Now we try to explicate this statement. First, we have to determine the types of expressions that appear in the above proposition. Alice is an agent, therefore of type ι. Ice cream (or rather the property of "being an ice cream") has the type $(o\iota)_{\tau\omega}$. We will use "σ" to denote this type. Thus certain entity (individual) instantiates the property of being and ice cream, or *icecreamness* if you will, in world w and time t if and only of it is an ice cream in that particular w and t. And finally, desire is a relation between an agent and certain entity that exhibits the desired properties. That gives us the type $(o\iota\sigma)_{\tau\omega}$.[6]

We have everything we need to offer an adequate analysis, which will take the following form:

$$\lambda w \lambda t.\mathsf{Desires}_{wt}\ \mathsf{Alice}\ \mathsf{Icecream}\ \ .$$

Statements of this type will be called *desire statements* and we will also use the letters δ_1, δ_2, ... to denote them. So by desires we will mean desire statements such as the one above.

Our next order of business is to give more precise account of the dummy-desires. Remember that we have said that dummy-desires are basically just "names" for desires, i.e., something that is not itself a desire but just mentions one. Some might have already noticed that the distinction between desires and dummy-desires distinction very naturally corresponds the the notions of using or mentioning constructions. And it is precisely this similarity that we will exploit to formalise the idea of dummy-desires.

Let's demonstrate this at following argument which is a variation of the one presented in our case study:[7]

$$\frac{\begin{array}{c}\text{Alice desires ice cream.}\\ \text{Alice desires ice cream} \rightarrow \text{Alice eats ice cream.}\end{array}}{\text{Alice eats ice cream.}}$$

[6] This definition is equivalent to Tichý's definition of *willing* in [7].

[7] The second premise is justified by our general account that some x is desire iff x has the ability to trigger and action towards x. The symbol → can be read as "leads to". There is also, of course, quite a strong idealisation, because having a desire does not necessarily lead to its fulfilment.

In our case study we have tried to show that this is not – generally speaking – a correct argument. Even though Alice desires ice cream, and despite the fact that desires lead to action towards their fulfilment, there is no guarantee that Alice will start such action. There is always the possibility that, e.g., stronger $\mathcal{R}$-desire might swing in at the last moment and overweight her desire for ice cream. In other words, despite what the arguments says, Alice might choose differently. So what went wrong?

On closer inspection we should be able to recognise that we have already met with similar argument before, at the end of Section 3. The argument there failed because of improperly identify use/mention distinction. More specifically, because of the illegal move of treating ${}^0[+\ 5\ 7]$ as having the same meaning as $[+\ 5\ 7]$.

Notice that what happened here is practically the very same scenario. We have treated the desire statements from premises 1 and 2 as if they were the same, but they are not. The first one just mentions the desire, while the second one uses it. In other words, the analysis should look like this:

$$\frac{\begin{array}{c}{}^0[\lambda w \lambda t.\mathsf{Desires}_{wt}\ \mathsf{Alice}\ \mathsf{Icecream}]\\ \lambda w \lambda t.\mathsf{Desires}_{wt}\ \mathsf{Alice}\ \mathsf{Icecream} \rightarrow \lambda w \lambda t.\mathsf{Eats}_{wt}\ \mathsf{Alice}\ \mathsf{Icecream}\end{array}}{\lambda w \lambda t.\mathsf{Eats}_{wt}\ \mathsf{Alice}\ \mathsf{Icecream}}$$

The first premise thus represent dummy-desire, for it only mentions the desire. On the other hand, in the second premise occurs as antecedent a genuine desire, which will upon execution trigger an action, i.e., will lead to the consequent of the conditional.

Thus we have to carefully distinguish between desire, represented by the following desire statement:

$$\lambda w \lambda t.\mathsf{Desires}_{wt}\ \mathsf{Alice}\ \mathsf{Icecream}\ ,$$

and dummy-desire (i.e., mentioned desired, or in other words, trivialised desire), represented by the following statement:

$${}^0[\lambda w \lambda t.\mathsf{Desires}_{wt}\ \mathsf{Alice}\ \mathsf{Icecream}]\ ,$$

and while the former, on execution, triggers and action towards getting and eating ice cream, the latter just mentions this desire. Simply put, ${}^0\delta$ is not the same as δ. Statements of the general form ${}^0\delta$ will be called *dummy-desire statements.*

So to sum it up: The incorrect analysis of arguments containing desires statements is caused by not properly distinguishing between used and mentioned desire statements, i.e., between desires and dummy-desires. While the former upon execution give rise to action, dummy-desires just name the procedure that leads to that action. The main morale of the story is that we have to first adequately analyse the premises of argument if we want to reach correct conclusion.

In the last section will be sketched the basic mechanism behind humean machine, or more precisely, behind the mechanism responsible for the communication between $\mathcal{D}$-engine and $\mathcal{R}$-engine.

4.4 Humean Machine

Humean machine consists of two engines (i.e., $\mathcal{D}$-engine and $\mathcal{R}$-engine) and assessing unit that facilitates the interaction between these two engines (see Fig. 1).

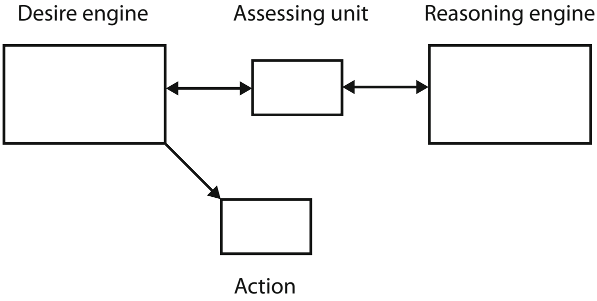

Fig. 1. Scheme of humean machine

The assessing unit has the following role: it checks if the dominant desire (i.e., the one with highest intensity) is the $\mathcal{R}$-desire (i.e., the "desire to act rationally"). If so, it calls $\mathcal{R}$-engine for advice on what to do and simultaneously sends it stock of "freshly" generated dummy-desires. After $\mathcal{R}$-engine comes up with a $\mathcal{R}$-conclusion, the assessing unit tries to find in $\mathcal{D}$-engine's desire base the desire counterpart to the $\mathcal{R}$-conclusion that the reasoning engine gave as a most reasonable course of action (basically trying to find a matching couple $\langle {}^0\delta : \delta \rangle$).

If this search is successful, assessing unit would send this information to $\mathcal{D}$-engine which would consequently execute the corresponding desire (i.e., δ) and thus triggering action towards the fulfilment of the $\mathcal{R}$-conclusion.

And if this search is unsuccessful, the $\mathcal{R}$-engine repeats the process until it is successful, i.e., until it provides such course of action that the agent is actually willing to do (meaning: finding such $\mathcal{R}$-conclusion whose desire counterpart can be found in $\mathcal{R}$-engine's desire base) or until the $\mathcal{R}$-desire seizes to be the dominant desire, i.e., until it is replaced by a new dominant desire. If this happens before feasible rational conclusion (i.e., such conclusion that is both conclusion of $\mathcal{R}$-engine and can be found in $\mathcal{D}$-engine's desire base) is reached, the desire engine takes over and simple does what new dominant (non-rational) desire demands.

This whole process can be roughly summarised in the following four steps:

1. If the dominant desire is $\mathcal{R}$-desire, call $\mathcal{R}$-engine and send it current stock of $\mathcal{D}$-engine's dummy-desires.
2. $\mathcal{R}$-engine reasons and returns possible rational course of action, i.e., $\mathcal{R}$-conclusion.
3. If there exists desire counterpart to $\mathcal{R}$-conclusion in $\mathcal{D}$-engine's current desire base (i.e., if for ${}^0\delta$ can be found corresponding δ), do what $\mathcal{R}$-conclusion suggests.

4. If there is none, repeat step 2 until 3 or until the dominant desire is no longer $\mathcal{R}$-desire.

These 4 steps actually capture very simple idea: keep thinking until you come up with conclusion you are actually willing to obey or until you get bored, and in that case, desires take over your decision making process.

The question of precise implementation strategies for these engines, as well as choosing adequate searching mechanisms for their respected information bases and and selecting the most suitable calculus for deriving the $\mathcal{R}$-conclusions, is left open for further research.

References

1. Montague, R.: Formal Philosophy: Selected Papers of Richard Montague. Yale University Press, New Haven (1974)
2. Duží, M.: Til as the logic of communication in a multi-agent system. Research in Computing Science 33(1), 27–40 (2008)
3. Hume, D.: A Treatise of Human Nature. Clarendon Press (1896)
4. Thomson, J.J.: Killing, letting die, and the trolley problem. The Monist 59(2), 204–217 (1976)
5. Tichý, P.: The foundations of Frege's logic. Forum Jazz Rock Pop. W. de Gruyter (1988)
6. Duží, M., Jespersen, B., Materna, P.: Procedural Semantics for Hyperintensional Logic: Foundations and Applications of Transparent Intensional Logic. Logic, epistemology and the unity of science. Springer, Netherlands (2010)
7. Oddie, G., Tichý, P.: The logic of ability, freedom and responsibility. Studia Logica 41(2-3), 227–248 (1982)

A Computational Behavior Model for Life-Like Intelligent Agents

Mohammadreza Alidoust[1] and Modjtaba Rouhani[2]

[1] Islamic Azad University – Science and Research Branch, Tehran, Iran
m.alidoust@hotmail.com
[2] Ferdowsi University, Mashhad, Iran
m.rouhani@ieee.org

Abstract. In this paper a novel computational behavior model is proposed which has a simple structure and also includes some of the major affecting parameters to the decision making process such as the agent's emotions, personality, intelligence level and physical situation. The effect of these parameters has been studied and the model has been simulated in a goal-achieving scenario for four agents with different characteristics. Simulation results show that the behavior of these intelligent agents are natural and believable and suggest that this model can be used as the decision making and behavior control unit of future life-like intelligent agents.

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

1 Introduction

Behavior Modeling is a very challenging aspect of research in the fields of artificial intelligence, control, sociology, psychiatry, psychology, economy, military, computer games, etc. and if performed correctly, we can improve the abilities of artificial agents and we can build life-like and social agents which can speak, think and behave like us. 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 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.

J. Romportl et al. (eds.), *Beyond Artificial Intelligence,*
Topics in Intelligent Engineering and Informatics 9, DOI: 10.1007/978-3-319-09668-1_12

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 models makes it difficult to validate them against the real human decisions [1].

In this paper a novel computational behavior model is proposed which involves a decision making strategy and some of the major influencing parameters in decision making process which make the intelligent agent more natural and believable. Another novelty of this paper is that it utilizes a simple structure that any other affecting parameters such as agent's memory can be easily augmented to in the future. The proposed model was tested on some agents with different intelligence and personalities in a goal reaching scenario. The aim of the intelligent agent in this scenario is to reach to its goal with minimum energy consumption and maximum enemy encounter prevention. The rest of this paper is organized as follows; In section 2 the emotion model and the basics of decision making strategy structure used in this paper will be described. In section 3 the role of influencing parameters that affect an agent's decision making process will be discussed. In section 4 the simulated results are depicted and section 5 discusses about conclusions and future works.

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.

In the following sections the above mentioned decision making strategy will be described and studied in a goal reaching scenario for an intelligent agent. Also the effect of influencing parameters in an agent's decision making process will be studied and augmented to the model.

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:

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 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.

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

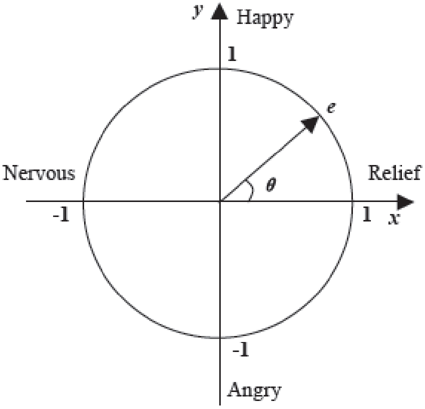

Fig. 1. Proposed Emotion Space of the Intelligent Agent

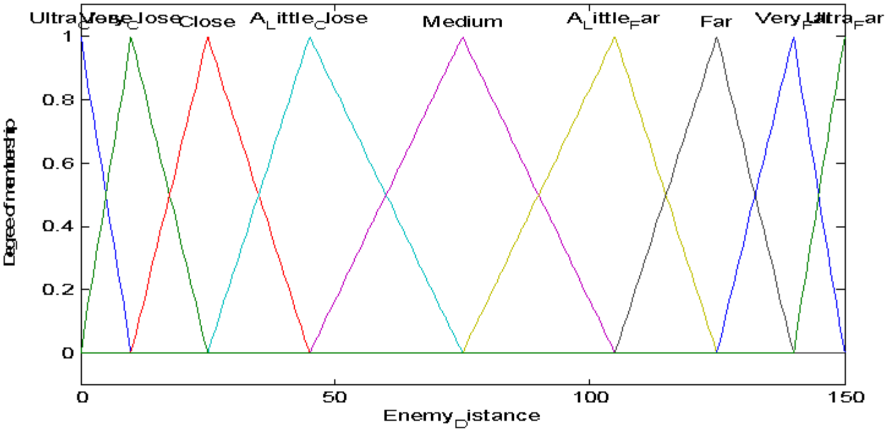

Fig. 2. Membership functions for input variable Enemy Distance

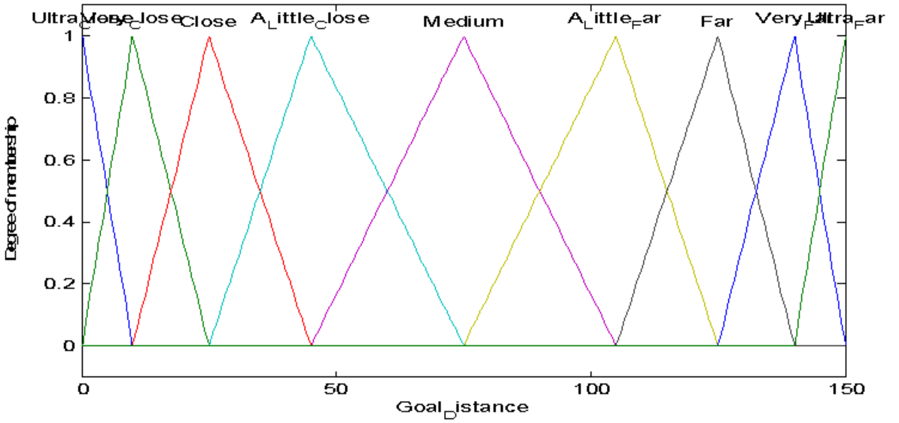

Fig. 3. Membership functions for input variable Goal Distance

Output Variables

Emotional Intensity trajectories x and y in Cartesian emotion space which both have 9 membership functions (UL, VL, L, AL, M, AH, H, VH and UH) equally partitioning the output space ranging from -1 to 1 that one of them is illustrated in Fig. 4.

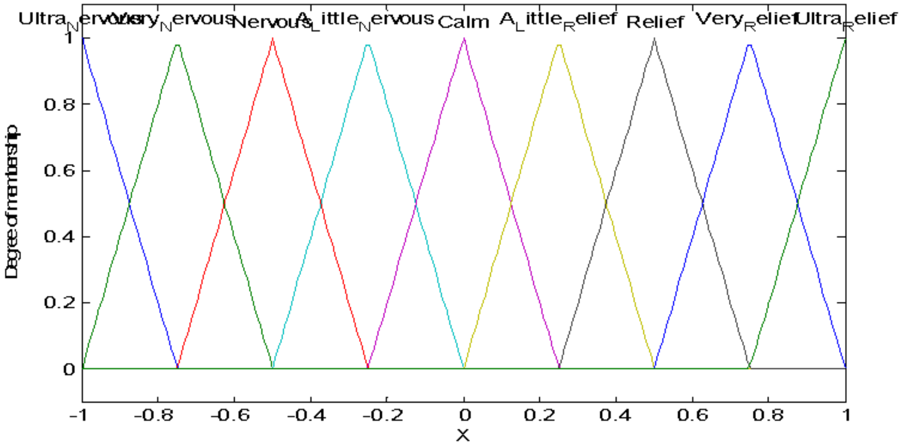

Fig. 4. Membership functions for output variables x and y

Fuzzy Rule Base

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

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

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 coordination to a circular coordination.

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

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

Which 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 $\underline{e} = (0.5, 30°)$ 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

2.4 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.

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;

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

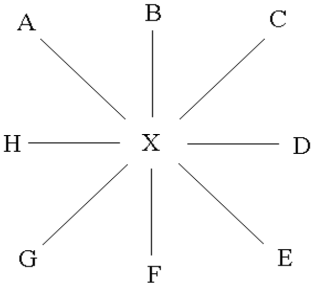

Fig. 5. Agent's possible movement alternatives

Where 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:

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

Which 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;

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

3 The Role of Moral, Substantial and Bodily Characteristics and Situations

The decision making strategy proposed by (12) leads to a deterministic and optimal agent behavior if e_c, g_c and α are considered static factors 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, physical situation, personality, etc. The mentioned characteristics play an important role in an intelligent agent's decision making process. For instance, it is intuitively 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. In addition, the decisions made by different people in the same situation are different due to their personality differences. The same condition is true when comparing the decisions made by smart and unintelligent people and also true when comparing the decisions made by tired and energetic people. This means the behavior of intelligent agents are to some extent stochastic rather than being completely optimal and deterministic. Therefore, this can be easily concluded that these characteristics are a major cause of the *natural* behavior of living agents. So, we have to add these characteristics to our raw decision making strategy defined by (12) in order to observe natural decisions from the agent. This goal can be achieved by dynamic factors e_c, g_c, α and index i. This will lead to more believable, intelligent and natural agents.

3.1 The Role of Emotions

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;

$$\begin{cases} e_c = 0.5 \\ g_c = 0.5 \end{cases} \tag{15}$$

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 probability equation as follows;

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

Here j is the time index and p^{i^*} is the probability of choosing the optimal solution and x^j is the emotion intensity's x-axis trajectory of current emotional state. Regarding (17) if the agent's emotional state is not "Nervous" ($x^j \geqslant 0$) the probability of choosing the optimal solution is 100%, and if its emotional state is "Very Nervous" ($x^j = -1$), the probability is 73.11%. So in this situation the agent may choose a wrong alternative and get hunt by the chasing enemy.

3.2 The Role of Intelligence

Generally higher intelligence causes identification of more possible number of alternatives. For example an intelligent chess player can forecast more possible future moves for himself/herself and his/her opponent. In the proposed scenario of this paper, the number of alternatives that an agent examines for its future move ($\{i\}$) is considered as an index of its intelligence. This means regarding to Fig. 5 a *moderate intelligence* agent examines 9 possible alternatives, while regarding to Fig. 6 a *high intelligence* (smart) agent examines 42 possible alternatives.

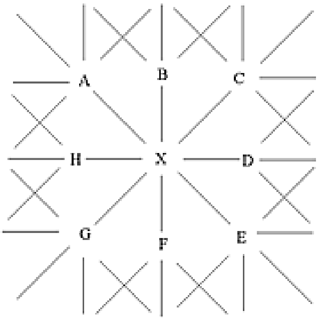

Fig. 6. A smart agent's possible movement alternatives

3.3 The Role of Personality

In this paper we assume just two opposite personalities for our agents, the *lazy agent* and the *energetic agent*. A lazy agent, has less tendency to move than an energetic one. Therefore, a lazy agent tends to stay in its current coordination rather than moving and consuming energy. The same as above, an energetic agent tends to move rather than staying and saving energy. Agent's personality can be enclosed by defining the initial value of α factor. α is the energy saving importance factor. So;

$$\begin{cases} \alpha^0 \geq \gamma & \text{, if personality} = \text{lazy} \\ \alpha^0 < \gamma & \text{, if personality} = \text{energetic} \end{cases} \tag{18}$$

In (18) γ is the *energy saving importance threshold* and can be identified based on the mass and other parameters of the agent. The time index 0 indicates the initial value of α.

3.4 The Role of Physical Situations

An agent's physical situations play a major role in its concentration and decision making process. An agent which is tired and has lost most of its energy resources cannot perform its decisions. Therefore, it is vital for a *tired* agent to consider energy cost of every alternative and disregard the alternatives which consume much amount of energy. This could be achieved by increasing α factor when the agent is tired and its energy is below a *tiredness threshold* (λ).

$$\begin{cases} \alpha(j+1) \geq \alpha(j) & \text{, if } e_k < \lambda \\ \alpha(j+1) = \alpha(j) & \text{, if } e_k \geq \lambda \end{cases} \tag{19}$$

Here j is the time index and λ is the tiredness threshold. When the agent's energy is higher than the threshold no change would be made to the value of α, but when it drops below the threshold, the value of α will increase. This will increase the agent's tendency to choose power-saving alternatives. The values of these parameters are arbitrary and can be tuned based on the requirements of the problem.

By adding all these moral, bodily and substantial characteristics, 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. 7. The stimulus sensed from the environment by the agent is interpreted by the EEFS. The resulting emotions and their intensities enter decision making block to choose the best alternative. Meanwhile, other affecting parameters such as the agent's personality, intelligence and physical situation affect this procedure. Finally if the emotion of the agent is "Nervous" the best alternative chosen by the decision making block will be altered by the quasi-Boltzmann stress block.

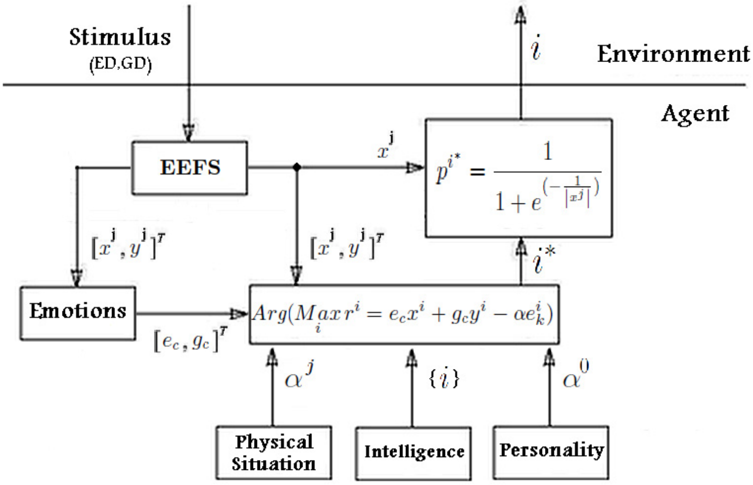

Fig. 7. Block diagram of the agent's decision making structure

4 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 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 or stay at their current positions.

Four examples of simulated behavior of some agents with different intelligence and personality are shown in Fig. 8–11. The Circle represents the location of the Goal; the Triangle represents the starting point of the enemy; the Square represent the starting point of the agent; Yellow points represent the enemy path when the agent is not in its eyesight (Enemy Distance $>$ 30 m); Magenta points represent the agent path when it is feeling "Very Nervous" (Enemy Distance $<$ 18.5 m) 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 (25 m), is greater than the eyesight of the agent when it feels "Very Nervous" and starts escaping from the enemy (18.5 m).

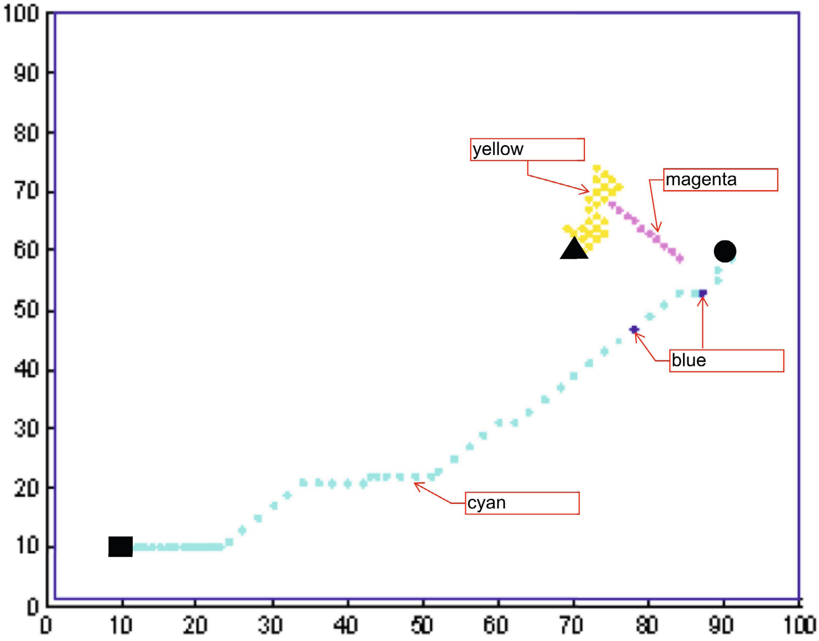

Fig. 8. Behavior of a lazy agent with high intelligence

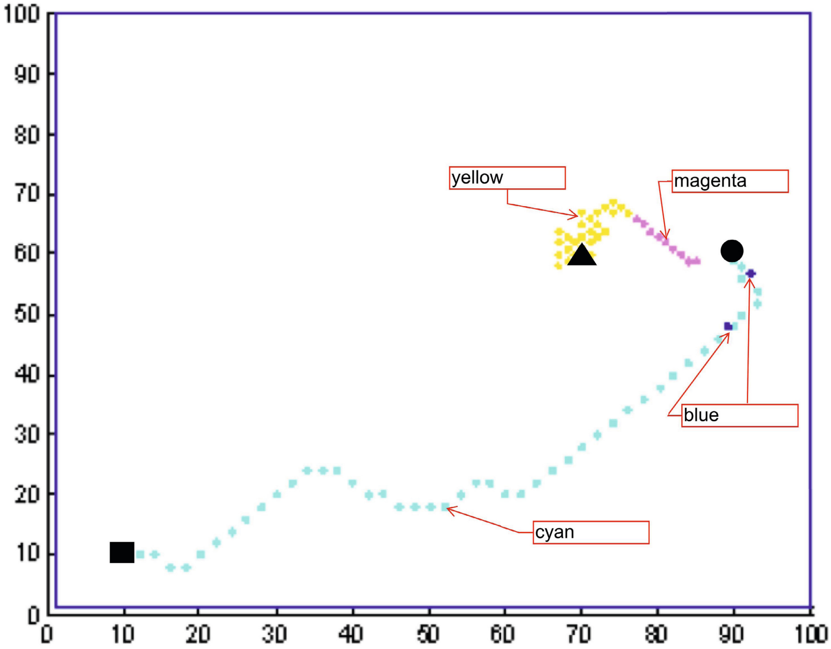

Fig. 9. Behavior of an energetic agent with high intelligence

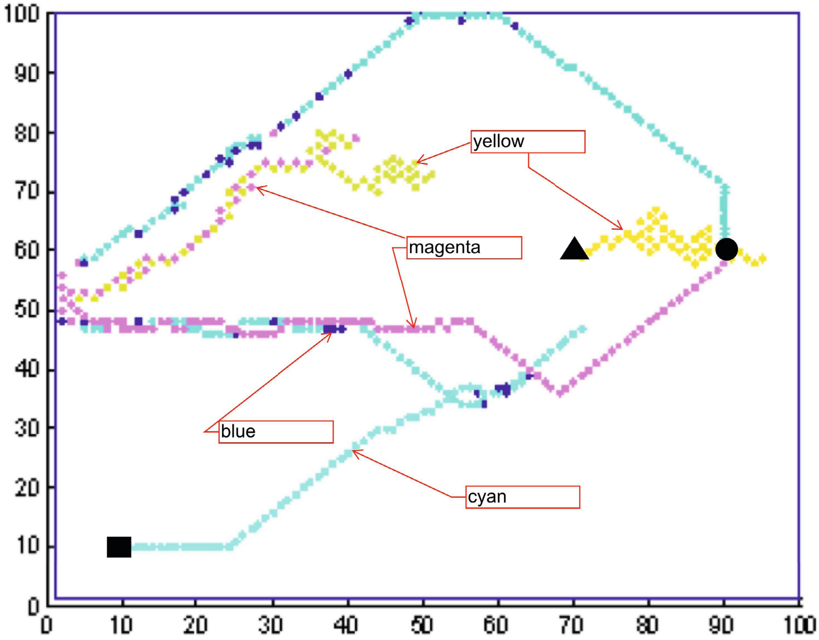

Fig. 10. Behavior of an energetic agent with moderate intelligence

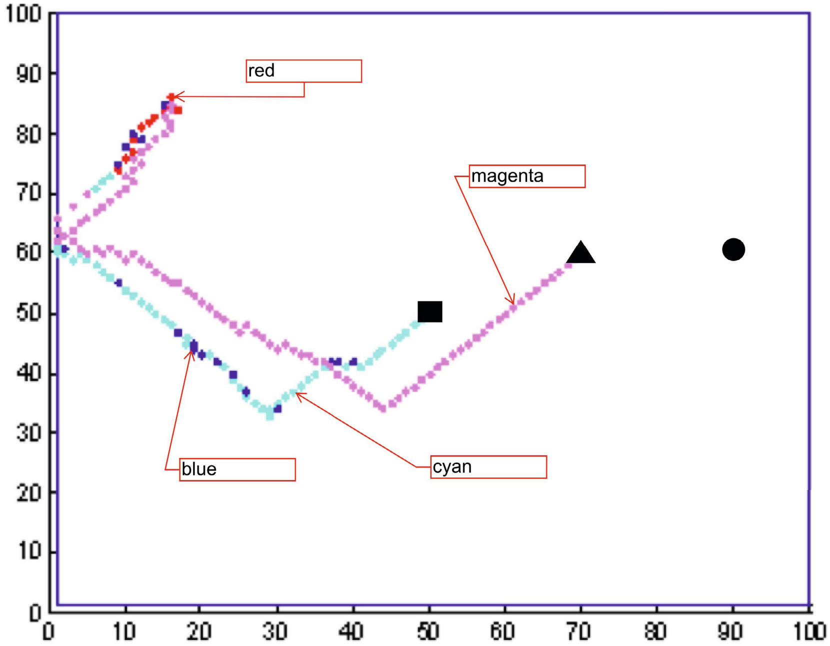

Fig. 11. Behavior of a lazy agent with moderate intelligence

Table 3. Average reaching to goal time for 50 trials

Agent Type	Average reaching to goal time	Rank
High intelligence, Energetic	53.2	1
High intelligence, Lazy	65.8	2
Moderate intelligence, Energetic	86.7	3
Moderate intelligence, Lazy	112.5	4

Table 4. Average remainder energy storage for 50 trials

Agent Type	Average remainder energy storage	Rank
High intelligence, Lazy	73.8%	1
Moderate intelligence, Lazy	60.6%	2
High intelligence, Energetic	59.0%	3
Moderate intelligence, Energetic	46.1%	4

5 Conclusion and Future Work

In this paper a novel model of behavior for intelligent agents was introduced and its validity was examined for agents with different personalities and intelligence level in a goal-approaching scenario. Regarding Fig. 8–11, the agent's behavior in this scenario is intelligent, natural and believable. Comparison results listed in tables 3 and 4 for different agent types prove this. Also the effect of the parameters discussed in section 3 on the agent's behavior was obvious. For example, as can be seen in Fig. 11, the agent with lazy personality and moderate intelligence faced lack of energy and also made a wrong decision because of its "Nervous" emotional state and so got hunt by its chasing enemy at coordination (17, 86).

The proposed decision making strategy is based on four basic emotions, but any other emotions can be augmented to the model easily. Augmenting more bodily, substantial and moral characteristics to the model can be easily achieved too. This leads to creating more natural and life-like agents. So, this model can be used as the decision making and behavior control unit of future life-like intelligent agents. Besides, the above mentioned characteristics play as the intrinsic dynamics of the agent so that they affect the agent's behavior (output). So, if these characteristics are sufficiently augmented to the model in terms of unknown variables, the behavior of any agent with any degree of complexity could be efficiently anticipated. This can be done by identifying these variables by applying sufficiently enough observations from the agent's previous behavior. Here we take a systematic look to the agent and consider it as a black box that its parameters must be identified.

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 soft computation methods such as Artificial Neural Networks, Reinforcement Learning and Evolutionary Algorithms could be the horizons for the future works.

References

1. Lee, S., Son, Y.: Integrated human decision mmaking model under belief-desire-intention framework for crowd simulation. In: 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 (1990)
9. Rao, A., Georgeff, M.: 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 (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: Decision Modeling and Behavior in Uncertain and Complex Environments, pp. 123–138 (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 (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.: 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., 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.: 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 (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)

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, artificial intelligence.

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 recognise 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 discriminative 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 endeavours.

J. Romportl et al. (eds.), *Beyond Artificial Intelligence,*
Topics in Intelligent Engineering and Informatics 9, DOI: 10.1007/978-3-319-09668-1_13

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 behaviour is defined as behaviour 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 colour.

Exception of the rule: A stone may not be placed on an intersection, if all adjacent points are taken by the opposite colour. (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 colour where every stone has at least one adjacent neighbouring 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 neighbouring stones of one colour. (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 summarises 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 sympathise 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 behaviour 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. Dodo Press (2009)
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 (2012), `http://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. In: Workshop on Representations and Processes in Vision and Natural Language, ECAI, pp. 78–80 (1996)
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. IFIP, vol. 135, pp. 159–174. Springer, Boston (2003)

9. Kocsis, L., Szepesvári, C.: Bandit based Monte-Carlo planning. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) ECML 2006. LNCS (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. Bouzy, B.: Backward history of Indigo (2012), `http://www.math-info.univ-paris5.fr/~bouzy/IndigoHistory.html`
12. Scarff, A.: AANNS explained (2012), `http://www.britgo.org/files/scarff/AANNS_Explained.pdf`

Answering Curious Questions about Artificial 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 plausible answers to the following frequently asked questions about artificial intelligence: (i) where knowledge comes from?; (ii) what is the "computational power" of artificial cognitive systems?; (iii) are there "levels" of intelligence?; (iv) what is the position of human intelligence w.r.t. the "levels" of intelligence?; (v) is there a general mechanism of intelligence?; (vi) can "fully-fledged" body-less intelligence exist?; (vii) can there exist a sentient cloud? (viii) how can new knowledge be generated? The answer to the first and the last question stems from the novel view of computation which is seen as a knowledge generating process. For the remaining questions we give qualified arguments suggesting that within the large class of computational models of cognitive systems the answers 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 following eight questions from the domain of artificial intelligence, all motivated more or less by curiosity: (i) where knowledge comes from?; (ii) what is the "computational power" of artificial cognitive systems?; (iii) are there "levels" of intelligence?; (iv) what is the position of human intelligence w.r.t. the "levels" of intelligence?; (v) is there a general mechanism of intelligence?; (vi) can "fully-fledged" body-less intelligence exist? and, last but not least, (vii) can there exist a sentient cloud? (viii) how can new knowledge be generated?

Undoubtedly, these are interesting questions to which qualified answers can only be obtained within the framework of the contemporary theories and views of computing and of cognitive systems. The mere fact that we are able to answer the above mentioned questions indicates that the underlying theories are really quite

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

J. Romportl et al. (eds.), *Beyond Artificial Intelligence,*
Topics in Intelligent Engineering and Informatics 9, DOI: 10.1007/978-3-319-09668-1_14

matured. We will see that when looking for the respective answers, our quest will be based on very recent results in epistemology and theory of computer science, indeed. The respective results concern a novel view of computation or non-standard computational models of cognitive systems.

The new view of computation is based on the recent work [1] where computation is seen as a knowledge generating process. Such an approach differs from the classical approach which sees computation as a process transforming information. The new approach concentrates to the main purpose of computation – i.e., knowledge generation – which presents the basis of intelligence. The non-standard computational models of cognitive systems used in the sequel cover a truly large class of systems. They present an important tool for investigation of cognitive systems since until now no cognitive mechanisms among natural cognitive systems (living organisms) have been identified that could not be modelled computationally. Our arguments will be based on four computational models each of which captures a different aspect of computational cognitive systems. In all cases, the answers are based on the recent work co-authored by the present author.

2 Answering the Questions

In order to give qualified answers to our questions we will refer to the recent results from philosophy of computation and to various non-standard computational (or algorithmic) models of general computational or specific cognitive systems.

The first and the last question concerning the origin of knowledge will be answered by referring to the recent idea that defines computation as knowledge generation process. The remaining answers will refer to various models of non-standard computations. While general computational models are suitable for answering very broad questions concerning the “power of AI” (questions (ii),(iii) and (iv)), answering a more specific question (v) and (vi) will need a fairly evolved model of an embodied cognitive agent with a specific internal structure. Question (vii) will be answered with the help of answers (v) and (vi) and of yet another unconventional model of general computations. Answer to question (viii) follows from (i) and (v).

2.1 Where Does Knowledge Come from?

Knowledge seems to be essential ingredient of intelligence: only knowledgeable agent can make the best of its intelligence. But – what is knowledge? What is the source of knowledge? How does an agent acquire it?

The questions related to the notion of knowledge are traditionally studied in epistemology which is the branch of philosophy concerned with the nature and scope of knowledge. Being a philosophical discipline, epistemology is more concerned with the definitions of knowledge, its characterisation and its relation to related notions such as truth, belief, and justification, and less in principles and mechanisms of knowledge acquisition and creation. Nevertheless, exactly the latter concern is central for understanding and designing knowledge processing algorithms which seem to be necessary for any artificial system displaying intelligence.

First of all – what is knowledge? It is an elusive notion which resists any generally accepted definition. If we are after a short definition, one of the shortest ones could be *"knowledge is facts, information, skills or behaviour enabling problem solving".* In Wikipedia, one can find a more extended definition:

> Knowledge is a familiarity with someone or something, which can include facts, information, descriptions, skills, or behaviour acquired through experience or education. It can refer to the theoretical or practical understanding of a subject. It can be implicit (as with practical skill or expertise) or explicit (as with the theoretical understanding of a subject); it can be more or less formal or systematic. [2]

Now – where knowledge comes from? In their recent paper, Wiedermann and van Leeuwen [1] have offered an interesting answer: *knowledge is the result of computation.* More precisely, they have coined a novel view of computation, seeing it as a process generating knowledge. In [1] the following thesis is proposed:

Thesis 1. *Computation is the process of knowledge generation.*

This thesis is supported by the evolution of application domains belonging to various type of computation. Roughly, the respective development starts with the classical Turing's acceptors and recognisers [3, 4], producing single bit of knowledge, proceeds via scientific computing delivering knowledge in the form of solutions of mathematical problems, further through operating systems, which generate knowledge controlling the behaviour of computer systems, and ends, so far, with the current search engines and question-answering systems delivering general encyclopaedic knowledge. The trend towards artificial general intelligence (AGI) systems capable to produce any human–like knowledge is clearly visible.

It is important to realise that a computation generates new knowledge based on the knowledge that is implicitly represented in the design of the computational system or is even explicitly stored within the knowledge base of such a system. Thus, one can say that knowledge generates knowledge.

It is advantageous to see knowledge contained in any computational system as a certain (more or less formalised) theory that is pertinent to a knowledge domain over which the system works and which is used by the systems in order to deliver its output.

If an agent can learn, then there are many ways for it to acquire knowledge: by reason and logic, by scientific method, by trial and error, by algorithm, by experience, by intuition, from authority, by listening to testimony and witness, by observation, by reading, from language, culture, tradition, conversation, etc.

The purpose of the *knowledge acquisition processes* is to discover new knowledge, enter it into the system and to order it into the knowledge already existing in the system. That is, in order the enable its later reuse new knowledge must be properly embedded into the existing theory representing an agent's current knowledge. Hence, any knowledge acquisition process builds and updates the existing epistemic theories. In this sense, knowledge acquisition is also a process of knowledge creation within, or 'inside' the respective computation. This again can only be done via computation.

We conclude with the answer that knowledge comes from computation.

2.2 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 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 [5, 6] 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. [7]. Thus, the answer to the second 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.

2.3 Are There "Levels" of Intelligence?

For answering this question we will 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. [8]).

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 "behaviours" (translations) can be realised. 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. [9]) 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.

2.4 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 [11] 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

[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. [10]).

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. [12]).

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 [12] – so called *red-green Turing machines*. This model characterises the second level of Arithmetical Hierarchy in terms of a machine model.

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 recognised if and only if on the inputs from that language the machine computations "stabilise" 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 2. *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. [13]).

2.5 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, randomised, quantum) each of which is characterised by a class of computationally equivalent mechanisms. We believe that this is also the case of cognitive systems which are but specialised 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 characterise the mechanism of intelligent systems as any computational mechanism generating the class of computations (resulting further into behaviours) that those systems are capable to produce or utilise. 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. [13]) (cf. Fig. 1) which is conformed with the recent state of research in the domain of embodied cognitive systems.

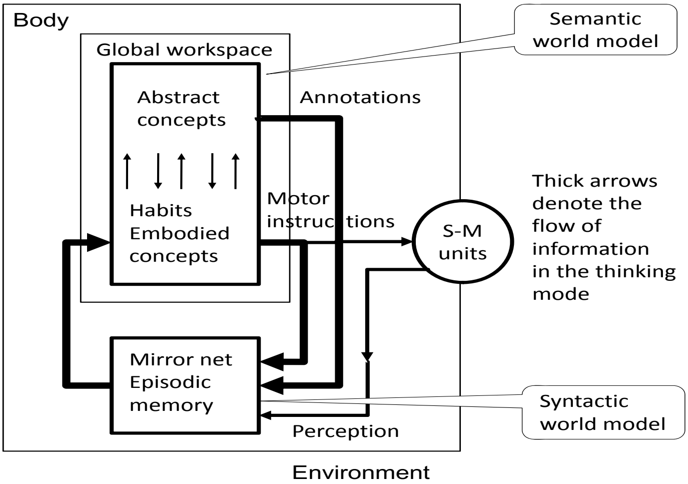

Fig. 1. The structure of a humanoid cognitive agent (HUGO)

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 realised 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 behaviour 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 [14], [15], and [13] 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.

2.6 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. [15]). 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, body-less 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. [16, 17]).

2.7 Can There Be a Sentient Cloud of Gas?

Written by by astrophysicist Sir Fred Hoyle the nowadays cult science fiction novel "The Black Cloud" [18] 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. [19]). 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 [20]. 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 [21].

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. [13]).

2.8 How Could New Knowledge Be Generated?

Essentially, the above mentioned question asks, whether an artificial cognitive system can be creative. A cautiously positive answer – which we are ready to offer – must at least indicate a constructive way how this is possible.

In Subsection 2.1. we have already mentioned that the purpose of the knowledge generation process, i.e., the purpose of any computation, is to produce new knowledge in reaction to the external or internal requests. But how is it possible for a computation to generate new knowledge that would not have been contained, in some way, in the initial data (read: in the knowledge base) of the computation at hand?

This is an interesting problem whose difficulty stems from the fact that known epistemological processes of knowledge generation are usually described as extrapolations of repeated observations, or of known facts, as some variants of an induction process. In this process, there is no creativity aspect: knowledge is merely transformed from one form to an other. This allows for no better explanation (or reasoning) than *"it has been so in the past, so it will similarly be in the future"*. However, it is reasonable to expect that the ability to create new knowledge must also include the ability to create new explanations of observed or conjectured facts which cannot be obtained by generalising the past experience or by putting the known facts together in some unexpected way.

So how could new explanations or conjectures be generated? One of the answers seems to be in the notion of analogy.

Analogy has been studied and discussed since classical antiquity by philosophers, linguists, scientists, lawyers and writers, and more recently also by cognitive scientists. The history of the subject is very rich. There are many definitions of analogy. For instance, *"analogy is reasoning or explaining from parallel cases"*; or *"analogy is a figure of language that expresses a set of like relations among* *two sets of terms"*. As an example, consider the analogy "*city* to *street* is like *country-side* to *river*".

What all these definitions have in common is a direct or indirect reference to natural language, to understanding, reasoning, explanations, and creativity. Within the theory of artificial cognitive systems all these notions are notoriously known as hard problems. Understanding of the underlying mechanisms evolves only slowly and therefore it is not surprising that the notion of analogy has seldom been approached from the viewpoint of requirements on the mental abilities of artificial cognitive agents.

One such a quest has recently been described in [22]. Here the author has shown the mechanism of analogy solving within the model of a humanoid cognitive agent described in Subsection 2.5. The proposed solution requires extensive searches over the agent's knowledge base that seek parallel semantic relationship among concepts entering into the analogy that are stored within the agent's semantic world model. Discovering of an analogy amounts to discovering of, in a sense, 'parallel' relationship between the concepts defining the analogy, or, in general, between two theories involving several concepts. This contributes to a better understanding of either theory since it enables to expect relations holding in one theory to also hold in its pendant theory. This is an important element of insight, explanation and understanding. Insight, understanding and explanation make only sense within a theory. They must follow from known facts or beliefs and rational thoughts. However, some theories can be based on incomplete facts or on wrong beliefs (cf. the flat earth theory). A discovery of semantic inconsistencies between alternative theories leads to a falsification of either theory. This seems to be the main source of new knowledge and thus, the main engine of progress (cf. [23]). Unfortunately, the respective mechanisms are so far poorly understood.

3 Conclusions

We have seen that using the recent novel view of computation, recent results from 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. [9]) 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 recent achievements of theoretical computer science, and especially, the theories of non-standard models of computations and the computational theory of cognitive systems that are seemingly unrelated go hand in hand in our quest for unraveling the secrets of intelligence.

References

1. Wiedermann, J., van Leeuwen, J.: Rethinking computations. In: Proc. of the 6th AISB Symposium on Computing and Philosophy: The Scandal of Computation — What is Computation?, pp. 6–10 (2013)
2. Wikipedia definition of knowledge (June 2013), http://en.wikipedia.org/wiki/Knowledge
3. Turing, A.M.: On computable numbers, with an application to the Entscheidungsproblem. Proc. London Math. Soc. 42(2), 230–265 (1936)
4. Turing, A.M.: On computable numbers, with an application to the Entscheidungsproblem: A correction. Proc. London Math. Soc. 43(2), 544–546 (1937)
5. van Leeuwen, J., Wiedermann, J.: The Turing machine paradigm in contemporary computing. In: Mathematics Unlimited - 2001 and Beyond
6. van Leeuwen, J., Wiedermann, J.: Beyond the Turing limit: Evolving interactive systems. In: Pacholski, L., Ružička, P. (eds.) SOFSEM 2001. LNCS, vol. 2234, pp. 90–109. Springer, Heidelberg (2001)
7. Wiedermann, J., van Leeuwen, J.: How we think of computing today. In: Beckmann, A., Dimitracopoulos, C., Löwe, B. (eds.) CiE 2008. LNCS, vol. 5028, pp. 579–593. Springer, Heidelberg (2008)
8. Verbaan, P., van Leeuwen, J., Wiedermann, J.: Complexity of evolving interactive systems. In: Karhumäki, J., Maurer, H., Păun, G., Rozenberg, G. (eds.) Theory Is Forever (Salomaa Festschrift). LNCS, vol. 3113, pp. 268–281. Springer, Heidelberg (2004)
9. Kurzweil, R.: The Singularity Is Near: When Humans Transcend Biology. The Viking Press (2005)
10. Cooper, S.B.: Computability Theory. Chapman and Hall/CRC (2003)
11. Penrose, R.: Shadows of the Mind: A Search for the Missing Science of Consciousness. Oxford University Press (1994)
12. van Leeuwen, J., Wiedermann, J.: Computation as an unbounded process. Theoretical Computer Science 429, 202–212 (2012)
13. Wiedermann, J.: A computability argument against superintelligence. Cognitive Computation 4(3), 236–245 (2012)
14. Wiedermann, J.: HUGO: A cognitive architecture with an incorporated world model. In: Proc. of the European Conference on Complex Systems, ECCS 2006 (2006)
15. Wiedermann, J.: A high level model of a conscious embodied agent. IJSSCI 2(3), 62–78 (2010)
16. Wiedermann, J.: Towards computational models of artificial cognitive systems that can, in principle, pass the turing test. In: Bieliková, M., Friedrich, G., Gottlob, G., Katzenbeisser, S., Turán, G. (eds.) SOFSEM 2012. LNCS, vol. 7147, pp. 44–63. Springer, Heidelberg (2012)
17. Wiedermann, J.: On the road to thinking machines: Insights and ideas. In: Cooper, S.B., Dawar, A., Löwe, B. (eds.) CiE 2012. LNCS, vol. 7318, pp. 733–744. Springer, Heidelberg (2012)
18. Hoyle, F.: The Black Cloud. Harper & Brothers (1957)
19. Wiedermann, J.: The many forms of amorphous computational systems. In: Zenil, H. (ed.) A Computable Universe: Understanding and Exploring Nature as Computation. World Scientific Publishing Co., Inc. (2012)

20. Wiedermann, J.: Nanomachine computing by quorum sensing. In: Kelemen, J., Kelemenová, A. (eds.) Păun Festschrift. LNCS, vol. 6610, pp. 203–215. Springer, Heidelberg (2011)
21. Tsytovich, V.N., Morfill, G.E., Fortov, V.E., Gusein-Zade, N.G., Klumov, B.A., Vladimirov, S.V.: From plasma crystals and helical structures towards inorganic living matter. New J. Phys. 9(8), 263 (2007)
22. Wiedermann, J.: The creativity mechanisms in embodied agents: An explanatory model. In: Proc. 2013 IEEE Symposium Series on Computational Intelligence (SSCI 2013), 2013 IEEE Symposium on Computational Intelligence for Human-like Intelligence (CIHLI), pp. 41–47 (2013)
23. Deutsch, D.: Creative blocks. Aeon (2012)

Biological and Artificial Machines

Vít Bartoš

Department of Philosophy, Technical University of 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. Differentiation between products of evolution and human-formed machines 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. But those assumptions are not in conflict with another assumption which is claiming that even biological entities are in fact kinds of computational machines.

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 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:

[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 consequence there is a selection on other levels); finally there is a heredity of favoring features.

J. Romportl et al. (eds.), *Beyond Artificial Intelligence,*
Topics in Intelligent Engineering and Informatics 9, DOI: 10.1007/978-3-319-09668-1_15

> 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?”

Some people think that this kind of question is inappropriate – a consequence of misunderstanding. You probably prima facie cannot see the link between something, which is usually understood as a abstract model of general computation (Turing machine) and between something completely different which is the process of evolution based on natural selection of individuals with differential fitness. But the point is in fact very simple. Edelman wants to emphasize that calculation executed on general computing machine (Turing machine) build up of discrete state system of transition could not be in any way compared with process of natural selection based on infinitely diverse “states” of evolutionary process on individuals. And the same principle according to Edelman's hypothesis we can apply on human thinking process where neuronal states are individualized as well.

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]

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. [3, 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.

2 Metaphysical Basis

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:

> 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. [4, p. 54, emphasized by the author]

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 Leibniz's 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):

> Thus the organic body of each living being is a kind of divine machine or natural automaton, which *infinitely surpasses all artificial automata.* For *a machine made by the skill of man is not a machine in each of its parts.* For instance, the tooth of a brass wheel has parts or fragments which for us are not artificial products, and which do not have the special characteristics of the machine, for they give no indication of the use for which the wheel was intended. But the *machines of nature, namely,*

> *living bodies, are still machines in their smallest parts ad infimtum.* It is this that constitutes the difference between nature and art, that is to say, between the divine art and ours. [5, pp. 254–255, emphasized by the author]

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.

3 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 ... [6, 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.

4 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.

5 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?
- 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 process – an evolutionary computation – are therefore the value systems of the organism whose task is multi-domain 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., Thompson, E., Rosch, E.: The Embodied Mind: Cognitive Science and Human Experience. MIT Press (1991)
2. Edelman, G.M.: Bright Air, Brilliant Fire: On The Matter Of The Mind. Basic Books (1992)
3. Turing, A.M.: Computing machinery and intelligence? In: Copeland, B.J. (ed.) The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life plus The Secrets of Enigma. Oxford University Press (2004)
4. Leibniz, G.W.: Theodicy. Open Court Publishing Company (1985)
5. Leibniz, G.W.: The Monadology and Other Philosophical Writings. Oxford University Press (1898)
6. Hillis, D.: The Pattern On The Stone: The Simple Ideas That Make Computers Work. Basic Books (1999)

Naturalness of Artificial Intelligence*

Jan Romportl[1,2]

[1] Department of Interdisciplinary Activities, New Technologies Research Centre
University of West Bohemia, Pilsen, Czech Republic
[2] Department of Cybernetics, Faculty of Applied Sciences
University of West Bohemia, Pilsen, Czech Republic
rompi@kky.zcu.cz

Abstract. This position paper offers an answer to the question about the difference between artificial and natural. By building up a dichotomy between physis and logos, it argues that this difference is given by language and by what can be grasped with words. It concludes with an assertion that Good Old-Fashioned AI (GOFAI) cannot create anything natural, whereas emergent AI can, because emergent phenomena are intrinsically natural, which is a very important fact for the AI field. The paper also offers a view on the difference between the roles of an AI engineer in GOFAI and in emergent AI.

Keywords: artificial, natural, intelligence, logos, physis, language, horizon, emergentism, GOFAI.

1 Introduction

Many AI lectures and textbooks start by discussing what *artificial intelligence* means. It is quite convenient – especially when introducing the field to young and keen students of engineering – to dismiss the problem by saying that AI is "*the science and engineering of making intelligent machines*"[1] and then continue with something "more useful", such as machine learning, artificial neural networks, pattern recognition, and so on.

A slightly different situation arises when AI is introduced to students of humanities, especially philosophy. They tend to split "artificial" from "intelligence", analyse both of them separately and then try to put them together in a sophisticated and holistic way. This second step almost never happens because the students – while in the first step deconstructing "artificial" – end in the void, concluding that the dichotomy artificial–natural makes no sense and that

[*] This work was supported by 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.

[1] In the words of John McCarthy, father of the term *artificial intelligence*, see http://www-formal.stanford.edu/jmc/whatisai/node1.html, or cf. Chapter 3 of this volume.

J. Romportl et al. (eds.), *Beyond Artificial Intelligence,*
Topics in Intelligent Engineering and Informatics 9, DOI: 10.1007/978-3-319-09668-1_16

there is nothing like *being artificial* or *being natural*. This is fair enough – it can happen to any term of natural language as long as it is attacked so strongly and deconstructed so thoroughly – but what if our world simply asks for a conceptualisation that distinguishes such an important categorical difference which can be seen for example between a plastic bag and a polypore, between city streets and a forest, between a house and a cave, or between an airplane and a bumblebee.

Maybe the artificial–natural dichotomy is not inherent to the world itself. However, it is surely inherent to our world conceptualisation, and above all, maintaining this dichotomy is simply useful, at least much more useful than dismissing it and pretending we do not see the difference (albeit the ability to see the difference does not imply the ability to say the difference).

I will now try to be rather constructive than deconstructive in this chapter and I will offer a possible way of characterising and formalising the distinction between artificial and natural, keeping in mind the usefulness that it should bear. My definition of natural and artificial should be useful for the AI discourse and should help when speaking about the disappearing human–machine divide seen as crossing (in both directions) the border between natural and artificial.

2 Human Naturalness

It is not very common to refer to ancient Greek philosophical terms in an engineering publication but here I feel such a step might be fruitful. The two concepts that I would like to bring to attention are *physis* and *logos*. In classical philosophy, they usually even do not form the opposites, but contrasting them here makes sense.

Physis is simply *nature* or *naturalness*, it is a concept referring to all things that grow on its own, intrinsically, to things that are in and created by the nature. It was used by the Greek god Hermes when he pulled out a plant to show Odysseus its intrinsic way of growth. *Logos*, on the other hand, means the whole complex and metaphorical concept of speech, meaning, human reason, rationality.

Physis connotates with raw untamed things, growth, procreation, reactivity, wilderness, warmth, dynamics, spontaneity, vortex, chaos, vagueness, mud, dust, rot, worms, turmoil, and also excessive growth (as in cancer). *Logos* connotates with clarity, purity, intentionality, geometry, logic, cold, statics, copying, algorithm, and also excessive loss (as in Alzheimer's disease).

If we put *physis* on one side of a continuous, somehow intuitive spectrum, and *logos* on the opposite side then we should put human right in the middle of the spectrum. Metaphorically speaking, human is a being of tension between *physis* and *logos*, a being producing and produced by this tension, a being that possesses about the same from both realms. I will try to explain this in more detail later. For now, it is important to note that "being as natural as human" thus means "having the same ability to balance in the equilibrium between *physis* and *logos*". Therefore, we would expect also a strong AI to be like this.

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 is a matter of *logos* and thus it gives the first blow to its pure naturalness. Tools of *logos* pull out pieces of inherently non-structured *physis* and construct shapes and objects from them. When Hermes showed the herb, drawing it from the ground to demonstrate its nature, its *physis*, the *physis* was already retreating. It was still somehow very strongly there, much more strongly than if Hermes had shown a plastic bag or a transistor, but no more in its pure form because artificiality has already crept in – the herb "being shown" is not the herb "being natural" inseparably in its *physis*.

What Hermes did, was something very typical for human, or even delimiting for human mind, thought and intelligence. If it were an animal instead of a human, it would not show/objectify/*name* the herb – it would simply transparently share with the herb their "unspoiled" and non-conceptualised *physis* together. On the other hand, if it were a machine instead of a human, it would operate only with purely symbolical representations completely detached from the intrinsic substance of the herb, and the herb itself would be substituted by a single symbol, or a symbolic representation of its geometrical model, or a symbolic representation of its molecular structure, or something similar. In other words, *logos* dissolves *physis*, and human is – in his nature – a steersman constantly oscillating around this unstable equilibrium where objects appear from the mud of *physis* before they disappear in the void of *logos*. We can also see this metaphor as a keen connotation to Wiener's and Ashby's cybernetics.

We can say that a major tool for such steering is *natural language*. Language in general is a long bridge between *physis* and *logos*, with deixis and protolanguages close to the bank of *physis*, formal languages, mathematics, geometry etc. close to the bank of *logos*, and natural language somewhere in between, where human minds operate. *Human naturalness* is thus something significantly different from *naturalness* seen merely as *physis* – human naturalness is an indivisible and intrinsic combination of natural and artificial, continuously re-enacted by the process of life itself. Therefore, the goal of the research field of Artificial (General) Intelligence is not building *physis* from *logos* (that would most likely be impossible) but rather pulling these two realms together and strike a new conscious mind on their frontier – this is probably doable.

3 Natural and Artificial Objects

When we use language to further analyse a freshly objectified (shown) object in more detail, we go step by step over this bridge and we start losing more and more of the object's naturalness. The object in itself stays the same but we receive more and more detached abstract concepts. Although such concepts are new objects transferable by means of language, they lose their connections to the inherent givenness of the original object. For example, no human being can describe in words how the root of a herb (or a cloud, a bird's nest, a coral) is *exactly*. The moment closest to the root's naturalness is when we *show* it (deixis), and since then the more we say about it in an attempt to fully describe it, the more artificial construct we get. At some point, the length of the description reaches beyond the limit of any human being and becomes manageable only by symbol-processing machines, having no meaning for human while being in this *logos* domain. For example, a "sentence" with 10 million "words" might be quite a good description of how the root is, but only as long as we interpret it as a 10-megapixel photograph of the root, forgetting everything about the language and humbly returning back to *showing* the root or at least its image, i.e. back to much more *physis*-related deixis. The *logos*-based interpretation of those 10 million symbols (i.e. reading and understanding them one by one) has absolutely no meaning for us.

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 in order to capture it by language.

On the other hand, *artificial* is imposed by language – artificial is that whose essence is fully determined by language. The artificial is a language abstraction drawn from the soil of *physis* and attracted by the clarity of *logos*.

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.

Apart from what has been said above about the artificial, we can add that the artificial is the means of our language-supported understanding of the world. However, not much more can be said about the artificial itself – the more we say about it, the more we feel that we are loosing its original concept; on the meta-level, the concept of artificiality itself defies being captured by language. Therefore, the concept of artificiality is very much natural itself – and so the artificial is the *natural* means of our understanding of the world.

4 Emergence and Artificially Built Naturalness

Through language, we can build a conceptualisation 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.[2] But – at least we know everything about this *logos*-originating scaffolding. We know everything about the world it encompasses, as much as we can know about a landscape from a map: the map is not for feasting one's eyes on the beautiful countryside, but for perfect orientation it is quite enough. The scaffolding itself is very much artificial and can be exemplified for example 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 naturalness – 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 (just like in case of the aforementioned 10-megapixel photograph). The machine will pile up the scaffolding floors on top of each other so quickly that it will soon reach the sky and even further. 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 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[3] 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 super-high 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.

[2] This is actually a rather poetic, informal and intuitive ultra-short introduction to *natural infinity*, one of the key concepts of Vopěnka's Alternative Set Theory [1, 2].

[3] Here 'finite' in the non-standard sense of Vopěnka's Alternative Set Theory, i.e. the opposite of *naturally infinite*.

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 [3] – 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, shaman, 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 artefacts. The disappearing of the human-machine divide – as Kevin Warwick in the first chapter of this book offers – is now much less surprising when we realise that the distance between *human* and *machine* is countless times smaller than between *physis* and *logos*.

References

1. Vopěnka, P.: Mathematics in the Alternative Set Theory. Teubner, Leipzig (1979)
2. Vopěnka, P.: The Great Illusion of 20th Century Mathematics and Its New Foundations. Preprint. University of West Bohemia, Pilsen (2012)
3. Haugeland, J.: Artificial Intelligence: The Very Idea. MIT Press, Cambridge (1985)

Index