AISB/IACAP World Congress 2012

Birmingham, UK, 2-6 July 2012

The 5th AISB Symposium on Computing and Philosophy: Computing, Philosophy and the Question of Bio-Machine Hybrids

J. M. Bishop and Y. J. Erden (Editors)

Part of

Published by
The Society for the Study of
Artificial Intelligence and
Simulation of Behaviour

http://www.aisb.org.uk

ISBN 978-1-908187-11-6

Contents

Foreword from the Congress Chairs

For the Turing year 2012, AISB (The Society for the Study of Artificial Intelligence and Simulation of Behaviour) and IACAP (The International Association for Computing and Philosophy) merged their annual symposia/conferences to form the AISB/IACAP World Congress. The congress took place 2–6 July 2012 at the University of Birmingham, UK.

The Congress was inspired by a desire to honour Alan Turing, and by the broad and deep significance of Turing's work to AI, the philosophical ramifications of computing, and philosophy and computing more generally. The Congress was one of the events forming the Alan Turing Year.

The Congress consisted mainly of a number of collocated Symposia on specific research areas, together with six invited Plenary Talks. All papers other than the Plenaries were given within Symposia. This format is perfect for encouraging new dialogue and collaboration both within and between research areas.

This volume forms the proceedings of one of the component symposia. We are most grateful to the organizers of the Symposium for their hard work in creating it, attracting papers, doing the necessary reviewing, defining an exciting programme for the symposium, and compiling this volume. We also thank them for their flexibility and patience concerning the complex matter of fitting all the symposia and other events into the Congress week.

John Barnden (Computer Science, University of Birmingham)
Programme Co-Chair and AISB Vice-Chair
Anthony Beavers (University of Evansville, Indiana, USA)
Programme Co-Chair and IACAP President
Manfred Kerber (Computer Science, University of Birmingham)
Local Arrangements Chair

The 5th AISB Symposium on Computing and Philosophy: Computing, Philosophy and the Question of Bio-Machine Hybrids

Turing's famous question 'can machines think?' raises parallel questions about what it means to say of us humans that *we* think. More broadly, what does it mean to say that we are thinking beings? In this way we can see that Turing's question about the potential of machines raises substantial questions about the nature of human identity. 'If', we might ask, 'intelligent human behaviour could be successfully imitated, then what is there about our flesh and blood embodiment that need be regarded as exclusively essential to either intelligence or human identity?' This and related questions come to the fore when we consider the way in which our involvement with and use of machines and technologies, as well as their involvement in us, is increasing and evolving. This is true of few more than those technologies that have a more intimate and developing role in our lives, such as implants and prosthetics (e.g. neuroprosthetics).

But the story of identity does not end with human implants and neuroprosthetics. In the last decade, huge strides have been made in 'animat' devices. These are robotic machines with both active biological and artificial (e.g. electronic, mechanical or robotic) components. Recently one of the organisers of this symposium, Slawomir Nasuto, in partnership with colleagues Victor Becerra, Kevin Warwick and Ben Whalley, developed an autonomous robot (an animat) controlled by cultures of living neural cells, which in turn were directly coupled to the robot's actuators and sensory inputs. This work raises the question of whether such 'animat' devices (devices, for example, with all the flexibility and insight of intelligent natural systems) are constrained by the limits (e.g. those of Turing Machines) identified in classical *a priori* arguments regarding standard 'computational systems'.

Both neuroprosthetic augmentation and animats may be considered as biotechnological hybrid systems. Although seemingly starting from very different sentient positions, the potential convergence in the relative amount and importance of biological and technological components in such systems raises the question of whether such convergence would be accompanied by a corresponding convergence of their respective teleological capacities; and what indeed the limits noted above could be.

In order to further explore these and related questions, the papers in this Symposium cover key related issues including, but not limited to: extended mind and extended cognition; brain simulation, physicalism, and pan-experientialism; the Chinese Room Argument and proof-theoretic justification; mathematical models of desire, need and attention; Cantor's diagonalization and Turing's cardinality paradox; bio-machine hybrids and cognitive technology; second-order cybernetics, autopoietic machines and structural determinism; computational creativity and swarm creativity; animats and bio-machine hybrids; Turing Machines and Gödel encoding; machine thought, identity, and

issues of recognition; human implants and prosthetics.

On behalf of the Organising Committee of this Fifth AISB Computing and Philosophy Symposium, we would like to thank all the members of the Programme Committee for their generous support, and for the excellent work in refereeing submissions. We hope that participants will find the event stimulating and enjoyable.

Mark Bishop (Dept. of Computing, Goldsmiths, University of London)
Symposium Chair and AISB Chair
Yasemin J. Erden (Philosophy, St Mary's University College)
Symposium Co-Chair and AISB Committee Member

Symposium Organising Committee:

Slawomir Nasuto (University of Reading, UK)
Kevin Magill (University of Wolverhampton, UK)

Symposium Programme Committee:

Paul Baxter (Plymouth University, UK)
Victor M. Becerra (University of Reading, UK)
Mark Bishop (Goldsmiths, University of London, UK)
Mark Coeckelbergh (University of Twente, NL)
Edoardo Datteri (University of Milano-Bicocca, IT)
Yasemin J. Erden (St Mary's University College, UK)
Tom Froese (The University of Tokyo, JP)
Matt James (BioCentre / St Mary's University College, UK)
Stanislao Lauria (Brunel University, UK)
Kevin Magill (University of Wolverhampton, UK)
Richard J. Mitchell (University of Reading, UK)
Slawomir Nasuto (University of Reading, UK)
Stephen Rainey (FUNDP, BE)
Ian Sillitoe (University of Wolverhampton, UK)
Porfírio Silva (Institute for Systems and Robotics Lisbon, PT)
Mark Sprevak (University of Edinburgh, UK)
Steve Torrance (University of Sussex, UK)

Hybrid Memory, Cognitive Technology and Self

Robert. W. Clowes[1]

Abstract. Recent years have seen an explosion in the production and use of technologies that allow us to record, store and recall ever-increasing amounts of information about our lives. Some welcome these trends as offering new possibilities for self-understanding and expression. Others think that things have already gone too far and worry deeply about what the future might hold. Does mem-tech really promise (or threaten) a radical change to the cognitive profile of human beings? If so, how are we to assess the possibilities and attempt to understand whether they offer a hopeful or dangerous turn in the human condition? This paper attempts to develop a balanced understanding of current trends in mem-tech and also consider some of its more probable future trends. In so doing it identifies four factors about the new memory devices: Capaciousness; incorporability; autonomy; and entanglement that suggest not just technical, but important psychological implications.

1 INTRODUCTION

Human nature and intelligence is not just a matter of our genetic endowment but relies heavily on a variety of factors including our cultural background, historically specific modes of thought and, not least, the pre-existing artefactual world into which we are born. Artefacts have in a variety of ways altered the lives of human beings and, directly or indirectly, the way we think.

Technologies which work more directly on our cognitive abilities we can call *cognitive technologies*[2].Yet, in fact, many developments of tools can bring with them changes in the modes or scope of human thinking. A favourite example of mine is cooking. Developing the ability to cook meat with fire may have dramatically reduced the amount of time early humans needed to spend finding food hence releasing time in which to think and, perhaps, invent culture. However, pragmatically including cooking as a cognitive technology makes the scope of any enquiry very large. We have to narrow this scope somehow.

Provisionally let's take cognitive technologies to be those technologies that perform functions which, were they to be performed by the human brain, would be regarded as cognitive[3]. No special claims are here made on whether or how cognitive technologies, or indeed other environmental resources might actually count as part of the mind. We will here side-step the ontological discussion around the extended mind and defend no strong position on whether cognitive technologies extend our minds [2, 3] or merely act as a new sort of environment in which they work [4]. Rather we are centrally interested in what happens to minds as they come to rely on the specific cognitive implications of digital technology, especially digital recording technology, handheld devices and all the paraphernalia of the mobile internet. As these technologies become increasingly pervasive in our culture, it is interesting to ask what if anything might be happening to our minds in the process. Most specifically we will focus on those digital technologies which may be reshaping human memory[4].

Despite avoiding the ontological question, we will use a terminology which suggests a tentative endorsement of the extended mind hypothesis by referring to E-Memory and O-Memory. The term O-Memory we here use to refer to organic or, perhaps better, organismic memory. O-Memory refers to an undoubtedly heterogeneous set of systems and processes which underlie the ways in which human beings and their brains retain and organise knowledge during episodes of experience which they can later bring to mind to put to work in a variety of ways. E-Memory similarly is used to refer to a heterogeneous bunch of devices and systems which fulfil similar functions either by replacement, extension or augmentation. One recent study[5] details how E-Memory[5] systems can support a range of human memory functions, including what the authors call the five Rs, namely: *recollecting*, *reminiscing*, *retrieving*, *reflecting and remembering intention;* the latter referring to way E-Memory systems (such as Microsoft Outlook) can allow us to track tasks, projects and actions that we intend to perform. Still, we should remember that the E / O Memory distinction is a conceptual division. One of the mains points of interest of this article is to shift our focus toward the current and future hybrid systems that are being forged as E and O-Memory systems interact in ever more intimate ways.

On the rough (and in several ways) problematic definition of cognitive technology just offered, we are spending ever-increasing amounts of time interacting with a new regime of cognitive technologies and especially E-Memory systems that have become a constant background to many everyday cognitive tasks. Google, Wikipedia and an ever-enlarging panoply of smart phones, personal gadgets, devices and software technologies, seem to be performing a variety of cognitive functions which either relate to, replace or augment O-Memory systems. These technologies include the encoding, storing and retrieval of memories and the full range of the five Rs just mentioned. And yet as these technologies and our habitual use of them is increasingly becoming a part of everyday life, the tendency is for

[1] Institute for Philosophy of Language, Universidade Nova de Lisboa, Av. de Berna, 26 - 4º piso, Portugal. Email: robert.clowes@gmail.com

[2] Richard Gregory coined the related term mind tools [1] to refer artefacts that have a direct effect on the way we think. However, the term cognitive technology seems to be in more general academic usage and even has its own journal: *Cognitive Technology Journal* published by Robert Rager Press.

[3] This way of looking at thing closely follows Clark and Chalmers original article on the extended mind. [2]

[4] Although we will mainly stay clearly of the ontological discussion it is interesting that memory seems to be becoming a crucial test instance for the extended mind.

[5] The authors were actually specifically discussing lifelogging, which we shall come to shortly.

them to become invisible, fading into the background of everyday life and skilled action.

Considering the amount of time we now spend interacting with these technologies, and arguably the possible profound implications for our minds, a series of important questions need to be addressed about what might be happening to us in the process. What are the cognitive implications of relying heavily on prosthetic technologies which fulfil tasks and functions that we once would have performed with our brains alone? To focus on memory, the main subject of this article: How are organic memory systems being changed by our encounters with and increasingly heavy usage of E-Memory devices? Given the central role memory plays in our cognitive architecture and its role in constituting our sense of self, is our use of E-Memory already or likely to start changing our basic cognitive profile? And, if so, how? Does this have implications for our broader humanity and the sorts of beings we are? These are deep questions then and difficult to answer - but we have to start somewhere.

In fact these questions have not passed entirely unnoticed in the wider culture; there are a series of authors who are deeply worried about what might be happening to us [6-11] in the process of our mass adoption of these technologies. Some of this work is a serious attempt to engage with what these technologies might be doing in interaction with our minds and some of it has a more sensationalist cast. This rather pessimistic outlook on what might be born out of this interaction between the mind and the new cognitive technologies is interesting in the light of some of the more utopian things that have previously been written about the internet's cognitive implications [12-14].

If anything, we are currently going through a backlash against such utopian thinking and so now, more than ever, we need to keep open the possibility that technology can add, as well as subtract, from the mind. Arguably, the history of technology and the mind up until now has been one where technologies with the most important intellectual implications, from writing, to the book, to the telescope, to the microscope have given the mind more than they have taken away. This article is an attempt to get a grasp on how mem-tech (digital memory technology) might already be having profound effects not just on organic (biological and traditional practices) of memory, but on our sense of self, and our wider processes of thinking.

2 E-MEMORY, LIFELOGGING AND ITS COGNITIVE IMPLICATIONS

Just as the amount and density of information that is being recorded about us in everyday life is ever-increasing [11, 15], so the ability of everyman to record the sound, images and many other sorts of digital traces of his life are showing a similar expansion [16]. The early twenty-first century has already seen a massive increase in the cheapness, availability and capacity of digital recording, storage and retrieval technologies that have placed an ever expanding arsenal of external memory technology in the hands of millions of people. The availability of cheap digital voice recorders and mega-pixel cameras embedded in mobile phones, as well as the powerful smart phones and tablets that many carry about all mean that increasing numbers of us are recording detailed records of our lives in ways which would have been scarcely possible only a few years ago. In addition, apps on smart phones and tablets are placing an arsenal of new software in people's hands that can put this information to innovative and exotic purposes.

The invention and widespread permeation of these technologies seem sure to have deep and widespread social consequences and perhaps offer to transform the way that both individuals and a society recollect and give meaning to both their personal and collective pasts. This process is continuing to the point that some now think it makes sense to believe that in the near-future we will seek to record the sum total of our experience: the dream (or chimera) of total capture [5] or total recall [17]. If there is little doubt that we have seen a technical E-memory revolution, then should we expect that our existing O-Memory systems will change and adapt to accommodate them?

Before tackling this question, however, it is worth asking whether what we are seeing is really novel. E-Memory is far from being the first technology to change how we use our organic systems. Arguably the history of the human race is in part of the history of how our O-Memory systems have been undergoing a constant process of elaboration and adaptation as we have created wave after wave of extended memory technologies [18, 19]. From spoken language – if it can be counted a technology [20] – through drawing and painting [21], to the development of counting systems, knots in rope, to writing systems [22, 23], through the development of record-keeping bureaucracies, the whole history of human art and technology can be seen as a history of revolutions in memory. And that is not even to make mention of techniques which have sought to reorganise (generally upgrade) human memory, from classical training in mnemotechics, to the medieval training use of memory palaces[23], to the rote learning systems practiced in twentieth-century schools. All of these inventions can be seen as important historical moments when our relationship with the technology of memory has undergone fundamental changes.

It is thus highly contestable that the purported reorganization of memory around particular technologies today is really historically unprecedented. Yet, it is surely worth pondering what, if anything, is new or distinctive about the particular cognitive technologies which are currently being developed. Only then can we decide if they might have novel cognitive and psychological implications for the human race. I suggest there are four aspects of the current crop of E-Memory technologies that have important qualitative or quantative differences from previous mem-tech and that we should focus our attention here to understand what is really new. They are:

1. Capaciousness & Comprehensiveness: E-Memory promises to record our everyday activities on a scale and with a fidelity and completeness that would have been practicably unimaginable under previous regimes of mem-tech.
2. Incorporability: E-Memory technologies potentially possess a transparency of use that makes them competitors (or complements) with certain of our internal resources. They are thus poised for deep and pervasive integration with O-Memory systems.
3. Autonomy: E-Memory repositories increasingly do not merely store data but actively process it. Thanks to tagging, indexing and AI systems we can expect E-Memory systems to not merely store and re-present information, but restructure it in a way that complements our native cognitive profile.
4. Entanglement –E-Memory often tracks interactions between people (or people and organisations). The form of the

data that composes many E-Memory stores is inherently relational[6].

Although there are no doubt many other dimensions of E-Memory technology which could have profound implications, each of the four I suggest picks out a quite fundamental aspect of the new mem-tech and, moreover, each is also a candidate for having important implications for O-memory, our minds more widely, our sense of self and even our humanity. We will now look in more detail at what is potentially novel about these aspects of the technologies before returning to their cognitive and psychological implications.

The most commented upon aspect of E-Memory is its promises to be able to record, and perhaps recall just about everything we might experience. This claim to a totality of capture and recall we have called **Capaciousness and Comprehensiveness**.

Perhaps the trend or idea that brings this out most clearly is lifelogging. Lifelogging consists of creating a personal and ever more detailed digital multimedia record of one's life as it happens. Compared to any previous technology to record memories, it makes an important departure: The aim of lifelogging is that rather than making the decision and effort to take a photo or record a telephone conversation, or make an entry in a diary; recording becomes effortless and the default setting[7].

The practice may be viewed as only making explicit a trend which is already deeply embedded among heavy users of the new digital technologies. Perhaps the most thoroughgoing and pervasive experiment so far attempted has been carried out by Gordon Bell and Jim Gemmel. Bell is a septuagenarian researcher with Microsoft but was an early pioneer of the networked computer. The project, directed by Bell and his Microsoft colleague Jim Gemmell, is called *MyLifeBits*[8]. As Bell tells the story, the project began with his desire to digitise, store and catalogue the books and articles he had written over the years. But, as the project progressed, Bell was no longer content with simply backing-up hardcopy but, as the technologies came online, Bell's aspiration became the creating of a digital record of everything he hears, thinks and sees. With this new orientation the *MyLifeBits* project turned its focus to capturing the ongoing stream of sensory information more or less as Bell himself received it.

Today, Bell not only has software on his computer to record and capture his every webpage visit, but he wears a SenseCam: a device which can be set to detect the presence of faces and was automatically set by Bell to take photos of those he encounters as he goes through his day[24]. Bell has also been experimenting to do similar things with audio technology and has equipment which records and attempts to categorise all of his conversations (and not just those on the phone). Bell now speaks about his aim as nothing less than to use electronic memory technologies to make a total record of an individual's sensory experiences: total capture [17]. In fact he more usually speaks about total recall: the ability to use all of this information to recollect any event in his past with total fidelity. Bell sees his quest for total capture and recall as in the tradition of inscription found at the entrance to the Oracle of Delphi: Know Thyself. Moreover Bell sees *MyLifeBits* as allowing him to develop new ways of knowing oneself that are a historic departure for the human race. Bell thinks his devices can allow him to know himself in ways no human has achieved before.

Viktor Mayer- Schönberger is another who believes that the possibilities of E-Memory and 'recording as default setting' portend profound effects on us, but he is far less sanguine about the prospects and, at the very least, he thinks it forces us to confront a new problem: How to forget,

> "through millennia, forgetting has remained just a bit easier and cheaper than remembering. How much we remembered and how much we forgot changed over time, with tools and devices emerging to aid our memory. But, fundamentally, we remembered what we somehow perceived as important enough to expend that extra bit of effort on, and forgot most of the rest. Until recently, the fact that remembering has always been at least a little bit harder that forgetting helped us humans avoid the fundamental question of whether we would like to remember everything forever if we could. Not anymore" [15] pg. 49.

Mayer- Schönberger believes we are on the cusp of changing a fundamental feature of our psychological lives with E-Memory technology. He worries that 'total capture', rather than putting us in deeper touch with ourselves, might reshape and even undermine our sense of self in profound ways. Much of this turns not so much on how much information we might store, but how we are starting to use it. (We shall return to this issue in section 4 below).

Our second factor, **Incorporability,** deals with the ways E-Memory might facilitate, bond with, augment or replace O-Memory such that the technology becomes second nature to the user, or, to use a more technical term, *transparent-in-use*. The sense here derives ultimately from Heidegger's observation that when we use a piece of equipment with which we are skilfully familiar, we cease to notice it as an object in itself with its own properties and our attention instead flows toward the task at hand and object on which we are working**.** Many technologies, including, in Heidegger's example, the humble hammer, can become transparent to the skilled user in the relevant respect. But arguably there are aspects of how E-Memory systems might become transparent-in-use that are qualitatively new. To pose this as a question: What happens when knowledge technologies[9] become cognitively transparent in this way?

There are several technical innovations behind these knowledge-technologies but of central importance is the availability of high bandwidth mobile connections, powerful

---

6 The inspiration for this notion comes from data-entanglement, see: [16]

7 Work which foreshadows lifelogging can be traced back at least to the 1980s in the work of such pioneers as Steve Mann who was experimenting with using digital cameras to record his everyday activities. In 1994 Mann set about using a wireless webcam to record is daily life 24 / 7 for artistic, experimental and in part also political reasons: Mann's project was political in that he was seeking to invert trends toward the surveillance of public space with an ever-growing arsenal of CCTV cameras, he aimed to surveil the surveillers.

8 A detailed description of this project and Bell´s motivations can be found in: [17]

9 I am here using the idea of knowledge technologies in a different way from cognitive technologies. The idea is supposed to be more specific and is used to mean technologies with a role in propagation of knowledge. Many internet technologies are prime examples.

mobile devices, cloud computing and, centrally, internet search. This ubiquitous computing technology makes it possible for us to have constant access to huge amounts of data, and mobile data applications, that may already compete with the authority of our organismic resources. As these technologies become more mobile (effectively a constant in our lives), ever easier to interact with, while our skills in using them deepen, it is likely we will tend to rely on them – incorporate them in our cognitive world - to an ever-greater extent.

Could there ever come a point where it is just easier to rely on ambient (or even biologically grafted in) memory devices than our own native O-Memory resources? Consider an example now familiar to many millions of users: Google Search. The internet based technology for finding information has for some time been used by many office workers dozens of times a day. As these search applications are increasingly accessed by mobile devices, they are rapidly becoming a constant part of the epistemic backdrops of our lives. With Google Search it is often quicker and easier to find out facts we might otherwise remember using O-Memory. Consider the act of bringing to mind the first name of an artist whose name is on the tip of your tongue, say the drummer with a band you once loved but haven't thought about in years. In the recent past you might wrack your brains trying to recall the name or try to think of something else assuming it will come to you in a short while. Today for millions of users of desktop computers and mobile devices you might instead type what you remember into the Google search engine. (I just typed 'drummer roxy mudic', I meant to type 'drummer Roxy Music', but my inaccuracy doesn't matter as the answer 'Paul Thompson' comes back in 0.3 seconds.) Typing a search query now often seems easier and in some cases more accurate than relying on our native O-Memory systems. In such circumstances typing search queries (or speaking into iPhones), has already become an everyday part of the recollection process itself.

Deep incorporation will turn on several factors of our use of these technologies. Of importance here is not merely how easy it is to interact with facility and effortlessness with our E-Memory devices, but how available they are to be incorporated into the patterns of everyday activities and thinking. To put this another way, it is not merely how *transparent-in-use* they become to us, but how deeply we come to rely on them. Other issues of importance are: The constancy and reliability of the resources; the constancy of our reliance on them; and perhaps centrally, our trust in them.[10] It is likely that deep incorporability does not merely depend on bandwidth or ease of use but on how comfortable we become with the idea of relying on E-Memory systems to make important decisions in our lives. Factors that influence this trust are likely to depend heavily on the social and institutional landscape in which these technologies emerge.

When one wrote an entry in one's diary - even if one were using it in the way of Otto from Clark and Chalmer's famous thought experiment [2] – one might reasonably expect the record to remain the same when one next came to look at it. E-Memory technologies however, have an ever more active profile and anything recorded with current tech is likely to be able to be represented back to its user in any number of augmented ways. E-memory devices can increasingly be expected to have the capacity to reorganise and repurpose the information they present in ways that are increasingly open-ended and reconfigurable. E-Memory 'stores' are really active repositories which increasingly transform and augment what they hold. This activeness and **autonomy** of E-Memory technologies might turn out be their most distinctive characteristic. How we adjust cognitively and socially to this autonomy is likely to be key in our future relationship with E-Memory systems.

To elaborate further, it is not merely that Google is easy to use and returns information quickly but that it is itself an active memory. Google, by storing pointers to, and ratings of, the mass of information which is available through the internet can return a page rank on any search term in a fraction of a second. Its database of content is constantly updated but, more importantly, for us, so are the algorithms and processes that are used to find that information. Information is not passively retained by Google but - in the pursuit of its twin goals of being useful and turning a profit - it is constantly being sifted and sorted with ever more sophisticated techniques with information undergoing processing and augmentation in various ways. (This is not even mentioning projects such as Streetview where Google is also creating huge new databases from scratch and using this to augment the information it holds and points at).

Thanks to the relative autonomy and active processing nature of E-Memory we can expect that it will become ever more transparent in use; although it is likely to become at the same time more *opaque in its workings*. The implications of this are that we may use it with felicity but increasingly have less idea of how it works. It is not just that technologies like Google may be passing beyond our powers of easy analysis but that companies like Google, in order to protect their competitive advantage, will continue to try to obscure the deep working of their technology.

There is a partial equivalence here with our native organic systems, as most people do not understand the deep workings of their minds either. (It has been the job of scientific psychology to attempt to understand the principles of organic human memory and there remains much work to be done.) But the type of autonomy of E-Memory means that the user's relationship with it is likely to be very different to his relationship with his organic memory. The main reason is arguably nothing to do with the technology *per se* but that the companies who are building E-Memory systems are likely to have different interests from the users of the technology. This may ultimately be a limit on how our trust relationships with the new cognitive technologies develop and perhaps upon whether we should ever ontologically consider such technologies as a part of our extended mind.

The way that E-Memory is likely to be organised, at least in the short term, is as much around the interests of corporations making software as anything we decide. What is made visible to others may not be what we desire. The conditions under which information is made visible to us is often something of which we are not even aware. *Edgerank,* the algorithm which Facebook uses to present timelines to its users is not in the public domain (*de facto* cognitively impenetrable). Most users are not even aware that they do not see a large proportion of the updates of their 'friends'. It may even be that, given the large amount of information that flows through systems like Facebook, such selective presentation is necessary, but this surely also has ethical and cognitive implications, especially if these systems become deeply entwined with our minds.

The autonomy of E-Memory technology is perhaps the qualitative dimension which sets it most apart most from previous regimes of memory technology. Moreover, it is likely

[10] All issues which echo Clark and Chalmers Extended Mind Paper.

that ever more active and perhaps autonomous E-memory systems will become increasingly pervasive. However as this happens, we are likely to find others sampling our activities to find patterns just as often as systems working to sample it for ourselves. This brings us to our fourth issue: **Entanglement**.

The idea of memory entanglement is that much of the data we are creating now, and the systems that control it, operate in part to stimulate or replicate recollection (such as Facebook history), is so deeply entangled with the lives of others to the point that it cannot accurately be considered data about *individuals* at all. What systems like facebook really track, are patterns of interaction. Social Media has been the main driver of this trend, but as it has expanded to encompass much of the activity of the internet some of our most personal data is now not only not held by us, but is deeply entangled with that of others.

Data from entangled repositories is already used to occasion memory processes, either according to our own wishes or because some organisation has chosen to remind us of something for its own purposes. The lines of who owns what are morally (if not legally) very blurred. Some are deeply worried by this [15], although there is a case to be made that there is really nothing new here. It is, after all, not merely our digital traces but our lives that are necessarily entangled with the lives of others. With or without digital media this is unavoidable. The desire to withdraw ourselves from public entanglement might really be a flight from the very idea of engagement with others [10].

Moreover, the types of entanglement made available by social media are probably changing rapidly. Some people apparently now use Facebook in the way people might have used diaries in the past. But a social network diary must function for very different purposes and presumably plays a different role for the individual.

Considering entanglement together with the autonomy point just discussed raises interesting questions about the determinants of how social media might help us to remember and forget. Facebook's edgerank algorithm is not a passive memory of our interactions with others. To the extent that its workings are opaque to us – and in part this is the flipside to transparency in use – we are not even aware of the criteria by which it might help us recall certain interactions with others. The properties of future E-Memory / O-Memory hybrid systems are likely to turn heavily on these sorts of interactions.

3 FUSING ORGANIC AND E-MEMORY

Just as the central thought experiment to illustrate the idea of the extended mind in the original article[2] featured Otto, who suffered with Alzheimer's, some of the most suggestive work on E-Memory and O-Memory integration has involved those suffering from memory deficits. Deacon Patrick Jones for instance suffered from Traumatic Brain Injury, leaving him with anterograde amnesia (inability to acquire new long-term memories) and difficulties in making use of existing ones. Deacon Jones describes the profundity of some of the difficulties in the context of meeting his children: "When they walk through the door, I don't know whether they will be three or thirty, I just try to interact with them as I find them."[25]

Nevertheless Deacon Jones has made considerable inroads into overcoming at least some of his problems by using the note taking software EVERNOTE and mind mapping software CURIO on his computer and through his iPhone. Thanks to cloud computing this software and his data store is available to him whenever he needs it in his everyday life. He uses EVERNOTE as a sort of long-term prosthetic memory and CURIO as an extension to his working memory. Many cognitive tasks that would be done entirely internally by most people are now being handled by the Deacon with his remaining organic resources in interactions with the E-Memory systems organised through an iPod, Tablet or his home computer. His ability to make use of this complex to edit a blog and look after a ministry (he has become ordained *since* suffering the most serious aspects of memory loss) is impressive, even inspirational. Given the profundity of his O-Memory deficits, Deacon Patrick's ability to live his life in a positive manner is undoubtedly extraordinary. It also indicates some of the possibilities E-Memory systems have to be integrated in the life and mind of an agent.

Another use of E-Memory devices by someone suffering from memory impairment is reported by the developers of the SenseCam in their attempts to help a female patient known as Mrs. B. who has severe memory impairment following limbic encephalitis) [24, 26]. Mrs B and her husband use a sensecam to record the events of their everyday life as they happen and then use desktop computer software to 'recollect' these events together. Mrs B's capacity when using the sensecam and then reviewing playback with her husband is as high as 70% recall for significant events (when she and her husband used written records as a comparison it is as low as 44%) [26]. It should be noted that the way they seem to be using the camera is not 'record-everything-by-default' in true lifelogging fashion. Rather, they take photos in the more traditional manner when they see something worthy of recording. Also note that Mrs B and her husband are using E-Memory in a highly collaborative fashion in order to aid her recollection: they sit at a desktop computer and review together pictures taken over a day. Nevertheless the SenseCam seems to have had positive implications both for Mrs B's O-Memory systems and for her life with her husband.

Or, consider again Gordon Bell's *MyLifeBits* project. Implicitly a major aim of the project appears to be to build an E-Memory that supports certain sorts of memory decline through aging. One part of this is incorporating face-recognition software into Bell's setup that can, on a real-time basis, report the name and contextual information - such as the last time Bell met a given acquaintance or the contents of an email from them - as Bell meets them going about his everyday life. So where Bell might have otherwise forgotten a one-time colleague's name, or some important information about her, his good devices are able to give the appropriate cue just as he needs it.

The intensive use of E-Memory might then eventually get a foothold in the senior population or among those with O-Memory disabilities, as people start to use E-memory as a straightforward replacement for fading organic memory systems, or with those who have O-Memory deficits for other reasons.

But as these technologies get used more widely it is likely they will start to support a whole range of extended cognitive functions. Similar systems to Bell's could use the internet to prompt users with information the user may never have encountered before, perhaps instantaneously Googling an unfamiliar colleague and providing unknown information as though it were remembered. Thus E-Memory technology might quickly come to support other cognitive functions as much as simply replace existing resources. In this way, E-Memory

devices might quickly shade into cognitive augmentation devices.

We might worry about this rapid evolution but it is also worth reflecting that this may be the natural trajectory of all technologies as novel uses are continually found for inventions not necessarily intended by their creators. If this is right, the path to the future is created as replacement or support seamlessly transitions into augmentation.

The open-endedness of this possible cognitive transformation is a source of worry to many commentators. Some have suggested that, as we rely ever-more on digital prosthesis, our organic capacities are under threat of atrophying [7]. Others that our humanity itself might be undermined [10]. What are the implications for those with 'normal'[11] memory profiles for the widespread adoption and incorporation of E-memory systems into their cognitive ecology? Could the reliance on E-Memory foreshadow a decline in our organic memory systems in the general population?

A basic premise of the organisation of organic memory systems and the deployment of neural resources appears to be 'use it or lose it'. Think for example of how somatosensory cortex remaps itself when a limb is lost. It is possible, that at least with regards to certain domains of knowledge, we will start to be able to explicitly remember less with organic systems as we use E-Memory systems more intensively. But the integration of E and O systems may be more complex that a zero sum game. The *complementarity principle* [27] holds that we will adopt extended resources insofar as they complement our basic (organic) cognitive architecture. The idea is that ambient resources will be useful insofar as they provide functions which, rather than replace, *contrast* with the brain's native methods of cognition and representation. If this is right, one would expect us to make use of E-Memory insofar as it makes available resources that are new and different from our native organic (or otherwise already enhanced) memory resources. On this analysis it is precisely because E-Memory – like other memory resources of the past – is offering something that is different from our native abilities, that there is likelihood it will be incorporated.

Here consideration of the idea of the extended mind has some interesting implications. If what really matters about us is the course-grained functional profile of our minds then the distribution of our cognitive resources between internal and prosthetic systems might really not matter very much. This may be one way of relaxing about the implied disuse of organic memory systems if we come to rely ever more heavily on electronic prosthesis.

Another reason turns on more practical concerns of how we use these technologies. Consider the satellite-navigation devices that many of us now use in our cars. Now consider using one to navigate an unfamiliar city over a period of weeks. One could imagine that using the sat-nav in this way might prevent one ever coming to learn to pattern of the city. Yet this does not seem to be the case. Instead the sat-nav gives one the possibility to drive to a destination while knowing next to nothing about where one is or where one is going other than the destination address. However, using the device over a period of weeks gradually familiarises the driver with the pattern of the roads in the city to the point the driver develops a good practical understanding of its navigation. Eventually it is no longer necessary to use the device. Really this should be no surprise as our O-Memory systems do not just stop working because we employ E-Memory devices and the sorts of interactions that may take place in true complementarity are likely to be subtle and complex.

If this analysis is along the right lines then rather than simply trading E-Memory for O-Memory it makes more sense – especially within the broader history of mem-tech – to think of an ongoing dovetailing process where technological and organic systems fuse in the overall organisation of the agent in a way that need not imply any necessary diminishment. Shouldn't we then learn to stop worrying and love the new mem-tech?

4 PERSONAL IDENTITY, SUPER SELVES AND FORGETTING

The idea that memory might be the key to our sense of self is a longstanding one going back at least to Locke, who held that while it was consciousness that constituted the unity of persons and self, memory was the means of connecting consciousness over time. In the contemporary discussion, the idea of an extended (or narrative) self which can be unified over time is a clearly related notion and so memory continues to play an important role in what many theorists think makes us persons [28, 29]. Yet from Reid's response to Locke until today it has been widely accepted that human memory is a problematic and fallible medium with which to achieve unification[12], for it is widely agreed that neither memory, nor narrative, are able to reliably achieve self-identity over time.

The *MyLifeBits* project and its successors might give us pause for thought, however. Our forgoing discussion of cognitive and memory augmentation suggests an interesting possibility. E-Memory, when used, as an adjunct to O-Memory might help us better fulfil the conditions for unity over time. Perhaps, by being able to store and then recall episodes in his life he might otherwise have forgotten through E-Memory systems, Bell, or other E-Memory pioneers, could potentially achieve a level of unity that us un-augmented humans cannot. This suggests the possibility that future humans, making extensive use of very authoritative and densely incorporated E-Memory systems, might have or become *Super-Selves*: Human beings whose unity over time is supported and guaranteed by their deep incorporation of an extended regime of E-Memory technologies and devices.

However enhanced unity over time might, in several ways, be counterproductive. Imagine Fred's teenage years are extensively documented by technologies like Facebook and feature episodes that in later years he would rather forget. Unfortunately, the social media traces Fred has left behind him are proving more persistent than he would like. Part of the problem is they are entangled with the traces left by others. Photos he would sooner now delete do not merely exist in his profile, but in the profiles of his 'friends' and moreover now proliferate through other systems that have reproduced them. Such traces plausibly might continue to shape and influence his sense of himself; its ongoing persistence could even constrain his future and what he might become.

For related reasons, some [8, 10] have started to worry that this persistence of certain types of entangled E-Memory might have seriously detrimental effects on humans beings in general,

[11] Of course there is no implication here that aging is not normal.

[12] For a nice recent discussion of the issues at stake and especially how these relate to recent findings about O-Memory, see [30]

but in particular on identity formation among young adults [10]. For, if we assume that some experimentation is necessary for the development of a stable and developed personality, then perhaps the capaciousness and authority of E-Memory might indeed risk undermining something essential in the human character: our capacity to move on from the past. Thus, we may come to see certain types of E-Memory as more of a prison than a source of useful reflection. Some now belief we need to develop institutional devices that declares some sort of moratorium on the potentially total retention of E-Memory [10, 31][13].

Mayer-Schönberger goes further and argues that forgetting is an integral part of human memory which plays an essential role in our cognitive profile and what it means to be human [15]. As Schacter [32] and others have pointed out, recollection at least is a largely reconstructive process. Each time we access a memory it can, at a neural level, be understood as being recreated. Forgetting is in part a process where our minds selectively maintain that which is useful for them and (as Freud knew) suppress much that is inessential or unhelpful. Forgetting may not be a bug in human memory but part of what the self-regulatory architecture of our minds does in order to have selves at all, at least as we currently understand them. Arguably our identity as unique human beings arises not just out of what we remember but out of what we forget. On this analysis rather than creating a super-self, E-Memory supported remembrance might actually undermine our sense of self.

On the strongest interpretations, E-Memory Entanglement becomes a sort of dominating determinant of our sense of self [8]. Mayer-Schönberger believes that if we come to accept that E-Memory can challenge the authority of our organic systems then we are in danger of losing something crucial about what it is to have or be a self. In a basic sense, if E-Memory systems seem more authoritative than our organic resources, our sense of self might become something estranged from us or alien. Yet this seems to approach a contradiction. Surely if there is anything which we have authority over it is our sense of self. Could the deep incorporation of E-Memory lead to a possible outcome where the sense of self is not really our own anymore?

This discussion of the entanglement of E-Memory brings up some difficult problems about the very meaning of the term self. Namely, self is taken by many – especially those trained in sociology – as rather than being something private (a hidden essence, character or set of memories), as something public and interactive. A dominant influence on the contemporary discussion of how social network technologies might interact with our sense of self is the work of Goffman [33] who is taken to say that the self should be understood less as an inner essence and more a public mask or series of performances[14]. (In fact, Goffman maintains a distinction between self and mask which many of his followers tend to collapse.) But it must be remembered this notion of self is very different to the tradition begun by Locke. The very idea here is that selves are unified thinking things, not masks. Nevertheless if this inward sense of self is strongly influenced by public performance then the facilities that social network technologies make available seem likely to play a role in this.

Even with the fledging E-Memory technology of today we do not consider ourselves infallible and our remembrance is often open to revision, especially if we find that other people – or sources – remember or portray things differently. We already have to factor in the vagaries of memory into our lives. Perhaps in the future it will just be a little harder to indulge in certain outright fictions about ourselves.

But it is not clear why this should endanger our sense of self *per se*. E-Memory's *de facto* entanglement with others in many ways is continuous with how a sense of self is constructed in the past: i.e. through interactions with others. That Mayer-Schönberger assumes rather than demonstrates that our sense of self (or personal identity) might be undermined by E-Memory is largely to do with how individuals come into conflict with organisations that can now more readily access and store more information about us than we would wish[11]. The growing imbalances of power between individuals and the companies that hold ever increasing amounts of information about us is undoubtedly a problem [e.g., 11] but this is a rather separate issue from the determinants of our sense of self.

5 SELF-KNOWLEDGE, POINT OF VIEW AND THE DEEP COGNITIVE BACKGROUND

Do the limits of the organic processes of consciousness and O-Memory really exhaust all we might wish to know about ourselves? This seems unlikely. The potential uses of E-Memory devices precisely promises to make available, or make explicit, information about aspects of our lives and ourselves that otherwise would be hidden in the background. Whether we will all always be happy with the forms of self-knowledge this information makes available is certainly questionable. But our felicity is surely no criterion for what should count as knowledge.

We need to take a step back from questions of power imbalances – important though they are – and ask whether E-Memory might nevertheless offer us new resources to constructively reflect on ourselves. Gordon Bell has been engaged in a practical form of this project and, as we have seen, conceives of the *MyLifeBits* project as a Delphic investigation into self-knowledge. We need to take seriously the claim that we could come to reflect on and know ourselves in ways that only this technology could make available. Let us consider again the claim that E-Memory can deepen self-knowledge by paying attention to the four factors which we previously held seem likely to be of the greatest cognitive and psychological import: Comprehensiveness; Incorporability; Autonomy; and Entanglement. In addition we will consider whether our interaction with systems with these properties might alter the sorts of beings we are.

Let us first consider some objections: It could be argued that Bell's dream of achieving an enhanced (perhaps even total?) form of self-knowledge with *MyLifeBits* is premised on a mistake about what self-knowledge is. Bell may be collecting

[13] Although there is not space to fully do this point justice here I think we must remember that not all societies have had a moratorium on youthful memories. The teenage years, where this sort of experimentation often occurs, are a particularly 20th Century invention and there have been many societies in the history of the world that have been hostile to this sort of personal experimentation. This is not to say that such experimentation is not important and valuable to us but it seems to stretch the issue to make it something necessary for the development of a sense of self *per se*.

[14] It's highly questionable if this is even a coherent interpretation of Goffman, see [30] page 104 – 105.

and digitizing data about himself with an unprecedented comprehensiveness, but that does not make it *self*-knowledge.

One reason to suppose this is that the data and E-Memory systems that Bell has amassed do not really count either as part of him, or his memory. Insofar as the E-Memory data does not deeply interact with Bell's own O-Memory systems, (it remains inferentially chaste), this seems a reasonable point. However E-Memory systems that are both easily incorporable and autonomous might quickly override such concerns. (We shall look at an example that touches on this point in a moment.)

Another objection is that self-knowledge, is not *merely* knowledge about oneself, but is only a distinctive category insofar as it is really the agent's own knowledge. To put this another way, self-knowledge proper has to in addition belong to the agent or be integrated into the agent in such a way as it can be said to have the property of *mineness*. Of course this does not solve the problem as we now have to be clear about what it would mean for an E-Memory system or its contents to have this property. One possibility of what we should want to mean by mineness is that the system is deeply integrated into the agent itself, and / or forms part of the agent's perspective, or point of view. E-Memory systems might thus really 'belong' to the agent insofar as they are deeply integrated into his cognitive processes, or form integral parts of his viewpoint.

Even if this is right, it is interesting that it may not disqualify even some current uses of E-Memory systems such as those developed by Gordon Bell. Consider how the SenseCam hangs around Bell's neck all day automatically taking and storing images. The images taken with it are – in a very literal sense – from Bell's point of view. Arguably this is not however the relevant sense of the term, for while the SenseCam may record information from Bell's point of view, it does not form *part* of his point of view. This raises the question of how and whether an E-Memory system, or information produced by that system, could ever come to count as part of one's point of view.

The following discussion will attempt to make it apparent that it is the details of exactly how an E-Memory system is incorporated with our organic systems – essentially the functional profile of their interactions – which will really count here. A deeply incorporated and trusted E-Memory system could indeed be considered to form a proper part of an agent's viewpoint, and systems that meet these requirements are much closer than we might think.

Rather than continuing to consider these points in the abstract, let us now consider three scenarios where E-Memory tech gets progressively more embedded in an agent's cognitive profile. We will consider a slightly fictionalized version of Gordon Bell's MyLifeBits system for illustrative purposes.

Scenario 1 – Here, type 1 E-Memory systems primarily operate in a passive way continually recording information in good lifelogging fashion that can later be reviewed by the agent. The striking feature of such systems – compared to previous regimes of memory technology - is the comprehensiveness of what is being recorded and the ease with which this is done.

It might be thought that such systems have only minimal cognitive implications, yet they already make available content that might contribute to one's self-knowledge in virtue of making available information that would otherwise be inaccessible or absent. This is broadly how Bell uses the *MyLifeBits* system now, although it is already shading over into another system more like our second scenario.

Scenario 2 – Let us imagine a more advanced E-Memory system which is more active, autonomous and deeply incorporated than the previous version. Instead of waiting the agent to perform a search it continually prompts him when it notices something that might be useful. The system is active in helping to organise the agent's attention.

Such a system might integrate a SenseCam and similar recording devices that capture images every couple of seconds and other contextual traces as the agent goes about his everyday business. It would automatically store these traces in an active database where various algorithms tag and do further processing on them. Those traces could then to be contextually recalled when useful. (We have already discussed such a system: the one Bell uses to retrieve a colleague's name when they appear in view.)

This mark 2 system is, in addition, constantly on the look-out for images or other traces that contain persons or objects already tagged as interesting. It 'notices' the recurrence of such interesting material in the current sensory stream and cues the user through some reality augmentation equipment. Over a period of time, as the agent interacts with and felicitously deploys such a technology, it might become – like Google search today - second nature for him (and thus *transparent-in-use*).

An interesting implication here is that even images or other traces that are stored in the database, and that the agent never looks at or consciously reflects upon, may nevertheless play a role in his cognitive architecture. This is because those stored traces in aggregation can trigger processes that cue or bias what is presented back to the agent – acting more like an organic implicit memory. Thus, the invisible and only indirectly known contents of the database might start to influence the agent's cognitive profile. (Cognitive opacity might here go hand-in-hand with transparency-in-use).

Scenario 3 – In a final scenario, an E-Memory system mark 3 incorporates many and varied autonomous systems which are hooked into the internet.

This near-future E-Memory technology continually sifts one's personal cloud-based data of multimedia "memories", perhaps constituted of every photo we have ever taken, every recording of our conversations, every email, etc, etc, and cross-references them against the resources of the internet.

Such a system might quickly start to seem less an adjunct to our mind and more as though it were an actual part of it. Because of its transparent usage, and the agent's reliance on it, such a system might become not merely a bias, but deeply incorporated with the agent's systems of attention. This third scenario suggests that the more autonomous and agentive technology, that we are already starting to see with some of today's web-bots, might start to play a more active role in the organisation of our thoughts.

Still the fact that mark 3 systems might incorporate in an *ad hoc* manner unknown internet based resources suggests that there may be fundamental trust issues here which would always prevent the user from treating such systems as though they were really parts of one's own minds. However, standards of trust may differ. Deep integration might turn out to depend in part on the agent's credulity.

Consider a scenario sketched by Andy Clark [34, 35] where a mark 3 E-Memory system has started to radically change what we mean by, and how we think of, ourselves. In a thought experiment Clark describes a subscriber to the Mambo-Chicken

Bot, a web-bot of the near future which "has been learning about, and contributing to, [his] taste for the weird and exotic for three and a half decades, coming online when [he] was five and first fell in love with astrophysical oddities." [35, pp. 128-129] In the thought experiment the subject has just discovered the Mambo-Bot has been disabled for the last three months and connects this with his feeling flat and uninspired for a while.

The idea here is clear; the autonomous and deeply incorporated cognitive technologies of the near future may well contribute not only to our sense of self but what we are; and in ways that do not have clear precedents in previous regimes of cognitive technology.

What are we to make of such systems? Are we to treat them as parts of the agent's memory, or adjuncts? And insofar as the agent relies on the retrieval and contextual information systems made available by advanced E-Memory systems, are we to regard those systems as part of the agent himself? Partly constitutive of his sense of self?

We have already hinted that part of this may depend on the cognitive transparency of the E-Memory system. At least in the *MyLifeBits* system, as the algorithms were largely set up by Bell to do tasks he intends, they can be naturally seen as extending his cognitive economy. Moreover, insofar as Bell has built those systems, he is likely to have a good sense of how far he can trust, rely upon and even defer to them. Such properties may not be maintained intact if someone else, who knew little about its workings, used the systems. The cognitive opacity of such systems to the user might make us unwilling to count them as proper parts of our minds essentially because we do not know enough about them to trust them; or indeed know enough to know we should not trust them. (This raises interesting questions about the cognitive transparency of minds more generally which unfortunately go beyond the scope of this paper).

What of the future for human beings where such systems are a commonplace? Such a future is likely to include social-media and personal Mem-Tech composing important tools for structuring and reflecting on ourselves. But, it is the autonomous and active nature of current and near-future E-Memory technologies that portends the most interesting and radical implications for who and what we are. If you doubt such a vision is in play with some of the top technologists of our time, consider this 2009 statement by Google executive Eric Schmidt on where he sees search technology going:

> "In the case of individuals, it's the model where the sum of what Google does becomes the third part of your brain – you know, there's a left brain, a right brain and there's a third part where the collaborative intelligence that Google can help bring to you really helps you get through every day."

There is reason to doubt E-Memory will fatally undermine our sense of having or being a self. In part this is because in order for there to be a deep integration between E and O-Memory it is likely to work according to something like the principle of complementarity and as a part of an integrated agent. So even though the resources on which the mind might draw are wide there is little reason to suppose that such a wide mind will not continue to have a sense of self. Even deep incorporation of E-Memory does not obviously imply the loss of that sense,

However E-Memory pioneers are increasingly becoming hybrid agents incorporating tools and software as it proves useful and changing their cognitive profiles in the process.

While we have tried to sketch some of the contours of how these changes might take place, only future research and practice will reveal its reality. It may, however, quickly come to seem that E-Memory might not merely facilitate new forms of self-knowledge, but new sorts of selves. We should not underestimate the agency both of practitioners and theoreticians in deciding how E-Memory should bond with O-Memory.

We have seen that E-Memory holds open the promise of novel possibilities for complementing our organic and culturally derived memory resources. A deeper understanding of these technologies' novel qualities, potentialities and also the complex and sometime contradictory roles memory plays in human life can only help us put them to more humanistic ends and perhaps avoid some of the more egregious pitfalls. There is little doubt however that they will be playing a larger role in our lives and, perhaps, our minds.

REFERENCES

[1] Gregory, R.L., *Mind in Science: A history of Explanations in Psychology*. 1981, Cambridge, U.K: Cambridge University Press.

[2] Clark, A. and D. Chalmers, *The Extended Mind*. Analysis, 1998. **58**: p. 10-23.

[3] Clark, A., *Supersizing the Mind*. 2008, Oxford University Press.

[4] Rupert, R.D., *Challenges to the hypothesis of extended cognition*. Journal of Philosophy, 2004. **101**: p. 389-428.

[5] Sellen, A.J. and S. Whittaker, *Beyond total capture: a constructive critique of lifelogging*. Communications of the ACM, 2010. **53**(5): p. 70-77.

[6] Carr, N., *Is Google making us stupid?* Yearbook of the National Society for the Study of Education, 2008. **107**(2): p. 89-94.

[7] Carr, N., *The Shallows: How the internet is changing the way we think, read and remember*. 2010, London: Atlantic Books.

[8] Greenfield, S., *ID: The Quest for Identity in the 21 st Century*. 2008, London: Sceptre.

[9] Lanier, J., *You Are Not a Gadget: A Manifesto*. 2010, London, England: Allen Lane.

[10] Turkle, S., *Alone Together: Why We Expect More From Technology and Less from Each Other*. 2011, New York: Basic Books.

[11] Pariser, E., *The filter bubble: What the Internet is hiding from you*. 2011: Penguin.

[12] Tapscott, D., *Growing up digital*. Vol. 302. 1998: McGraw-Hill New York.

[13] Shirky, C., *Cognitive Surplus: Creativity and Generosity in a Connected Age*. 2010, London: Allen Lane, Penguin.

[14] Negroponte, N., *Being digital*. 1996: Vintage.

[15] Mayer-Schönberger, V., *Delete: The virtue of forgetting in the digital age*. 2011: Princeton Univ Pr.

[16] Gemmell, J. and G. Bell, *The E-memory revolution*. Library Journal, 2009. **134**(15): p. 20-23.

[17] Bell, C. and J. Gemmell, *Total recall: how the E-memory revolution will change everything*. 2009: Dutton.

[18] Vygotsky, L.S., *Mind in society: The development of higher psychological processes*. 1978, Cambridge Mass: Harvard University Press.

[19] Donald, M., *A Mind So Rare: The Evolution of Human Consciousness*. 2001, New York / London: W. W. Norton & Company.

[20] Donald, M., *Precis of the Origins of the Modern Mind: Three stages in the evolution of culture and cognition*. Behavioral and Brain Sciences, 1993. **16**: p. 737-791.

[21] Mithen, S., *The Prehistory of the Mind*. 1996: Thames Hudson.

[22] Ong, *Orality and Literacy: The Technologizing of the word*. 1982: Methuen.
[23] Olson, D., *The world on paper: The conceptual and cognitive implications of writing and reading*. 1994: Cambridge University Press.
[24] Hodges, S., et al., *SenseCam: A retrospective memory aid*. UbiComp 2006: Ubiquitous Computing, 2006: p. 177-193.
[25] Marcus, G., *What if HM had a Blackberry? Coping with amnesia, using modern technology*. 2008.
[26] Berry, E., et al., *The use of a wearable camera, SenseCam, as a pictorial diary to improve autobiographical memory in a patient with limbic encephalitis: A preliminary report*. Neuropsychological Rehabilitation, 2007. **17**(4-5): p. 582-601.
[27] Sutton, J., *Exograms and interdisciplinarity: history, the extended mind, and the civilizing process*, in *The Extended Mind, London: Ashgate*, R. Menary, Editor. 2006, Bradford Book, MIT Press: London, England. p. 189-225.
[28] Parfit, D., *Reasons and persons*. 1984: Oxford University Press.
[29] Gallagher, S., *Philosophical conceptions of the self: Implications for cognitive science*. Trends in Cognitive Sciences, 2000. **4**(1): p. 14-21.
[30] Baggini, J., *The Ego Trick*. 2011: Granta Books.
[31] Mayer-Schönberger, V., *Useful void: The art of forgetting in the age of ubiquitous computing*. 2007.
[32] Schacter, D.L. and J.T. Coyle, *Memory distortion: How minds, brains, and societies reconstruct the past*. 1997: Harvard Univ Pr.
[33] Goffman, E., *The presentation of self in everyday life*. Garden City, NY, 1959.
[34] Clark, A., *Dispersed Selves*. Leonardo Electronic Almanac, 2009. **16**(4-5).
[35] Clark, A., *Natural Born Cyborgs: Minds, Technologies and the Future of Human Intelligence*. 2003, New York: Oxford University Press.

I remember me: neuroprosthetics, memory and identity

Dr Yasemin J. Erden [1]

Abstract. The emerging interface between nanotechnology and information and communication technology (ICT) looks set to radically enhance the production of neural implants or neuroprosthetics. Responses to these developments, and specifically from within dominant functionalist accounts of mind, hold that neuroprosthetics are likely to prove no more problematic for concepts of human identity than, say an artificial pacemaker. This paper investigates claims of this nature and shows that such accounts rely on an over-simplified model of brain function.

1 INTRODUCTION

The belief that human persons are composite, made up of mind and body (even soul) continues to dominate contemporary Western values and beliefs. Our sense of collective and individual identity is deeply affected by how we perceive this composition. Yet some of these traditional views of the human person as composite (mind, brain, body, soul) are already being radically challenged by some areas of scientific research and technological innovation. As technology becomes more sophisticated, more invasive and yet more common within industrial societies it has the potential to influence conceptions of human identity substantially, as well as compound and engender new issues in human dignity.

ICT implants are one such innovation, and raise important associated conceptual and ethical questions about freedom, dignity, privacy, security, consent, trust, control and equity of access. For example, the emerging interface between nanotechnology and ICT for the development of brain/nervous system implants may lead to more sophisticated neuroprosthetics with systems of information storage and external retrieval, without a clear indication of how to manage this access both in terms of legislation and while taking into account issues of security and human dignity. This paper will propose that there are deep philosophical and ethical issues arising (or compounded) as a result of these developments, which have not yet been sufficiently addressed, particularly in relation to the mind/brain/body relationship.

Popular conceptions of the crossover between nanotechnology and computer science or robotics typically presume the construction of nanobots, or nanites. Yet these sorts of futuristic developments are controversial to say the least, and at the very best unlikely in the near to medium term. As such, the focus of this paper will instead be on current developments in electronic devices with dimensions as small as some bacteria (nano-enabled ICT implants), and especially those implanted in the brain or nervous system (neuroprosthetics). One example is a silicon chip ICT implant that could either restore or enhance memory by replacing a damaged hippocampus (which plays a key role in forming memories). Claims that devices of this sort could supposedly perform the same processes as the brain section it is replacing will be investigated alongside associated ethical issues.

Accordingly, this paper follows two related avenues. The first concerns the nanotechnological enabling and expansion of the ICT-implant concept by the use of nanomaterials and nano devices, while the second issue will be to consider the impact of these implants, primarily neuroprosthetics, for human identity and dignity. The study at this juncture will therefore be in two layers: the impact of ICT-implants in general, and the prospective impact of nano-enabled ICT-implants in particular. The primary goal will be to explore some of the deep philosophical and ethical issues arising or compounded as a result of these developments. Before we consider this, it is pertinent to ground our understanding of the mind/brain/body problem.

2 IDENTITY AND THE (DIS)EMBODIED MIND

The idea that a human person is composite, divided into either mind/body, body/soul, or mind/body/soul, permeates Western thinking at least as far back as Ancient Greek thinkers. This approach persists in contemporary popular discourse on identity, despite significant arguments against a division. Challenges include standard materialist and eliminative materialist accounts, for example, or functionalist interpretations that permeate the scientific medical view, alongside emerging neurodeterminist perspectives. In fact, there has been substantial empirical evidence against such simplistic divisions (such as from lesion studies, practical psychology and current neuroscientific research) for many decades, yet the sense of division remains tied to our narrative on identity (we still talk of the mind much more than we refer to the brain). Reasons for this are complex, and it is not my intention to tackle these details here. What is relevant to my purpose however is to focus on what sort of impact this division has had.

Certain Western approaches to the human person (philosophical, theological) have been informed by a belief that mind and/or reason are either independent of, or the governing force behind, the body. The body, in turn, is oftentimes considered to be inferior to the mind. This poor estimation has tended to manifest in how the body has sometimes been treated, valued and its needs explained and addressed. At worst the body is considered irrelevant, superfluous or something to be mastered. This latter approach in particular has had many patrons within philosophy, central examples being both Plato and Descartes. Both figures, in contrasting ways, contribute to the exclusion of the body as an essential aspect of our identity. For the former, the body represents a prison-like shell through which we are bound to a lowly, corporeal/worldly existence that ultimately disappoints our higher aspirations. This unfortunate physical existence distracts us from philosophical concerns. As a result it is but a burden for the philosopher, who is left 'despising the body and avoiding it, and endeavouring to become independent' [1, 65c-66e]. The separation of mind and body is

[1] St Mary's University College, London, UK
✉ erdenyj@smuc.ac.uk

furthered by Descartes who defines the body as a *machine*, one that can be doubted, distrusted, and generally ignored when compared with the value we find in the mind:

> And although I may, or rather, as I will shortly say, although I certainly do possess a body with which I am very closely conjoined; nevertheless, because, on the one hand, I have a clear and distinct idea of myself, in as far as I am only a thinking and unextended thing, and as, on the other hand, I possess a distinct idea of body, in as far as it is only an extended and unthinking thing, it is certain that I, [that is, my mind, by which I am what I am], is entirely and truly distinct from my body, and may exist without it. [2, p. 156]

Numerous examples of this sort could be cited, and despite certain efforts towards the dissolution of such divisions (as noted above), dualistic accounts of various types permeate popular everyday narrative about our fragmented identity. Body and soul, mind and body, mind/body/soul. Whichever way you look at it, division has defined our way of presenting, measuring and evaluating our human identity. Advances in emerging technologies look set to further challenge these perspectives, particularly as it is likely that technological innovation will increasingly permeate a greater proportion of our lives. With this rise there is an increased likelihood, for instance, and even in the short and medium term, of further cementing our identities as 'networked individuals' [3, p. 21]. As technology of this sort becomes more invasive and yet more common, it seems fair to say that this is likely to influence our conceptions of human identity not insubstantially. ICT Implants are one such example to think about in this respect.

Before we turn to these however, there is one theory of mind that requires attention, primarily because it has commanded so much attention over the past 15 years or so. This is Clark and Chalmers' extended mind concept. In their seminal paper The Extended Mind they pose the question 'Where does the mind stop and the rest of the world begin?', to which their response is that wherever the line is drawn, it is not predetermined by any physical demarcation of skin and skull [4, p. 10]. Instead they propose a theory of mind that involves what they call 'active externalism', a conclusions from which is that when I write my thoughts on paper this may be considered functionally the same as the thoughts as they exist in my mind. They defend this claim with the example of Otto, who suffers with memory loss, and Inga, who does not. In order to aid his memory, Otto writes things down in a notebook. In this example, the authors claim that Otto's notebook functions just like the memories in Inga's mind, since the contents of the notebook can guide action, in the same way as the beliefs in Inga's mind can guide action.

While there may be some overlaps with the discussion of memory and identity in this paper, it would be a mistake to assume that the issues at stake are therefore the same. To begin with, the extended mind theory of Clark and Chalmers' rests on a presupposition regarding what sort of memory each example represents. To explain this further we can consider De Preester's [5, p. 134] response to Clark and Chalmers' claims. She states that there is in fact an important difference between what Otto and Inga do, since 'Otto has to re-appropriate his belief each time he needs the belief at stake and looks it up in his notebook'. This is unlike Inga who instead has what De Preester calls 'explicit ownership' over the thought [5, p. 134]. I do not wish to enter the debate about 'ownership' regarding thoughts, but I think her overall point about a difference between *extension* versus *tool use* (where in her argument the notebook would be considered the latter) is an important one, and has ramifications for my argument here about identity. For this reason alone (although there are others) the extended mind theory has little impact on my central argument. My point is not that our standard human composition is the *only way* that our identity could be formed (whether now or in the future), nor do I wish to make any explicit claims about the nature of memory. Instead my argument is only to think about the sorts of *changes* we are making and their possible ramifications. In this respect Clark and Chalmers' argument offers a valuable example, since the invention of writing did in fact have a rather substantial impact on human identity and evolution. As have many other *tools* and through our evolution. My suggestion is that we ought to be particularly mindful about current and future developments that involve real *extension*, such as neuroprosthetics, and think ahead about what their possible impacts may be.

3 IMPLANTS

Artificial implants can serve important medical functions in humans and, at least historically, have tended to be passive medical devices. Artificial valves and joints are some of the more common examples. This area is expanding however, with the application of ICT and nanotechnologies, toward the development of more sophisticated, primarily *active*, medical implantable devices [6, pp. 119-120]. There are a number of significant ethical and philosophical issues arising from the latter category in particular, many of which require immediate attention.

In March 2005 the European Group on Ethics (EGE) presented 'Opinion No 20: Ethical Aspects of ICT Implants in the Human Body' to the European Commission. In this document [3, p. 24] they offer the following:

> Does a human being cease to be such a "being" in cases where some parts of his or her body –particularly the brain - are substituted and/or supplemented by ICT implants? Particularly as ICT implants can contribute to creating "networked persons" that are always connected and could be configured differently so that from time to time they can transmit and receive signals allowing movements, habits and contacts to be traced and defined. This is bound to affect their dignity.

One example that brings to the fore these sorts of conceptual and ethical issues is the recent development of an 'artificial hippocampus'. A vital region of the mammalian brain, the precise function of the hippocampus has sometimes proved difficult to pin down, yet there is some consensus that it has a pivotal role in memory formation:

> While historically there has been debate over the precise role of the hippocampus in various functions of the limbic system, it is now widely accepted that its major contribution lies in the formation of long-term memories and the process of learning. [7]

The hippocampus is often an early region to suffer damage from Alzheimer's Disease. The *brain prosthesis* would supposedly mimic hippocampus function, rather than simply stimulate brain activity, and as such the 'silicon chip implant will perform the same processes as the damaged part of the brain it is replacing' [8]. Whether the chip can achieve this aim remains to be seen, but in 2009 the research team leading this innovation, led by Theodore Berger, was awarded a 4-year $16.4 million DARPA grant to further their research into restoring lost memory function. This follows a hefty $24 million investment into similar research on Brain Computer Interface (BCI) programs, split between six different laboratories [6, p. 145]. In financial terms the potential to repair or enhance the brain is being taken seriously.

Yet a number of important questions arise here. For example, is the artificial hippocampus concept dependent on a mechanistic and reductionist notion of mind/brain, and what are the implications of this? What kinds of scientific *models* are available here, are they adequate, and what are the limitations of such modelling? How are traditional boundaries between health and sickness being challenged? Is it the case that people ought *never* to accept certain defects or illness, old age and death, and if not, what are the limits and what kind of limits are they? These are only a sample of questions; more will be considered below.

In fact, perceptions of the *person* as a 'work in progress', something to be fixed or upgraded permeate our popular culture as well as our approach to health and well-being. With the ideals and visions of leading-edge biomedical science and technology, there is a shift of questionable sustainability in our corporeal identity. Such challenges to human identity are likely to deepen as scientific advances lead to more sophisticated technologies. If technology becomes more invasive and yet more common within industrial cultures it is likely to influence both concepts as well as conceptions of human identity substantially.

Other pertinent developments in implant technology are targeted at mood control. These include implantable neurostimulation devices that can modify electrical nerve activity and in this way be used for the treatment of severe depression. Another relies on input-output interactions or BCI, which alongside *neurofeedback*, could allow user/computer interface through electrical impulses [9]. These mechanisms might also be used in future for the management of depression. While still in its infancy, these sorts of developments seem likely within the short to medium term. For example, in October 2003, neurobiologists led by Miguel Nocolesis reported success in teaching rhesus monkeys to consciously control a robot arm using their brain and visual feedback via 'a closed-loop brain–machine interface' [13, p. 193].

Innovations of this kind raise all sorts of ethical and conceptual issues, not least with regard to what counts as *severe* in terms of depression, and the methods for diagnosis. Symptoms similar to severe depression can of course be displayed in other circumstances, for instance in bereavement; yet, distinguishing between these two states is a problematic task. This is particularly because for diagnosis to be effective it requires, amongst other things, understanding and coherence on the part of the patient. Other questions about the nature, and (potentially) even the *value* of depression (why it may occur, what it does, how it feels to live with it) must not be ignored, particularly if the treatment is irreversible (the link between, e.g. depression and creativity has been debated for many years). Despite this, research into the possible impact of these technological developments for the nature of human identity is limited. Yet it cannot be denied that our identity is formed by a multitude of experiences, emotions, memories and so on. This includes those experiences or emotions that may cause us pain, such as depression or bereavement. While I would not wish to claim that those who are suffering ought to continue to do so, I remain sceptical of those who would simplify such conditions to the purely medical, and develop technologies accordingly. Particularly when their role in our lives and formation of identity remains so complex and uncertain. As Fiedeler and Krings [11, p. 1] rightly point out, there is sometimes 'a technological optimism', which is problematic, not least because of the tendency of some research to view the brain as 'just a complex but physico-chemical determined machine'.

To understand the ramifications of this question of identity let us return to the example of the artificial hippocampus. Since the hippocampus is key in the formation of memories, this artificial prosthesis could be used to both restore and enhance memory. What if the hippocampus replacement leads to an improved capability for producing more accurate memories? Would this amount to repair or enhancement? Would more accurate, more complete, or even just *more memory* actually be an improvement at all? Does this take into account a broader idea of the value in either scale or quality of those memories we typically form, whether consciously chosen or otherwise? Would an artificial hippocampus negate this potential for choice altogether? Again, my point here is not to say that the technology is flawed, but only that these questions need to be considered in some depth, and to query how often they are. For example, in a discussion regarding what sort of enhancements we might expect from converging technologies, Burger [12, pp. 167-168] lists 'better' senses, memory and imagination. All of which, as the above shows, are contentious and require further consideration, not least because the idea of *better* presupposes current boundaries to be somehow insufficient or limiting. What would it mean for our imagination to be better? And what might that mean for how we live our everyday lives?

In areas like nanotechnology and ICT, presuppositions about our identity inform what sort of contributions these technologies can make to our lives and well-being. Progress is presumed based upon what may in fact turn out to be significant misunderstandings regarding the complexity of identities and identity-formation. Another example of the leap from repair to enhancement is with *prosthetic cortical implant*s, which although originally developed to restore aspects of sight, could allow visually unimpaired people to access 'information from a computer based either on what a digital camera sees or based on an artificial "window" interface' [6, p. 147]. This is without even addressing other issues connected with the restoration of sight in the first place.[2]

The EGE Report [3, pp. 23-24] notes that legislation is needed 'in order to avoid a situation in which society is becoming more and more dependent on such intrusive technology in order to provide social security', yet it is unclear whether this legislation would be effective in stemming the tide of increasing technology dependence. Even within this report there remains uncertainty

[2] The controversy surrounding cochlear implants for children is a pertinent example of the controversy surrounding the 'normalisation' of children, often with insufficient attention to broader social, ethical and psychological issues.

about definitions. While they allow for 'medical purposes' and 'legitimate social applications' for ICT implants this does not escape key questions: what counts as repair (legitimate or otherwise), and what counts as enhancement? As they note, the *borderline* between repair and enhancement is by no means strict [3, p. 23].

It seems highly probably that our memories play an extremely important role in our identity, and it is clear that we do not (perhaps cannot) always understand the multifarious ways in which this occurs. Before we take for granted the benefits of technologically advanced implants for these purposes, we need to be clear about what this might mean for us. This is by no means to say that these developments might not prove valuable. Indeed, there is much evidence to the contrary. Those affected (directly or indirectly) by neurological disorders such as Alzheimer's disease and illnesses like depression may welcome these developments, and there is perhaps much to welcome. My point is only that we ought to consider the broader effects of these developments, and to proceed with caution. This is particularly true where nanotechnology is concerned.

4 CONVERGING TECHNOLOGIES

In broad terms it is clear that 'technology miniaturisation trends, such as smaller sizes, lower power consumption and increased performance' [13, p. 3186] will substantially affect the structure of implants over the coming years. How the developing ability to manipulate matter at the nanoscale will affect *specific* developments of ICT implants is, however, still uncertain.[3] Nanoscale devices, with at least one dimension less than one-tenth of the approximate diameter of a red blood cell, are now being developed, and some of these will probably have biomedical and neurological applications. For example, an engineered nano-fibre and a human neurone may be brought into functional contact.

The rise of nano-ICT interface in implant technology (by, for example, using nanofibres), as well as nanocomputing looks set to develop exponentially in the coming decades, and reflects a broader trend towards *convergent technologies*:

> The phrase 'convergent technologies' refers to the synergistic combination of four major 'NBIC' (nano-bio-info-cogno) provinces of science and technology, each of which is currently progressing at a rapid rate: (a) nanoscience and nanotechnology; (b) biotechnology and biomedicine, including genetic engineering; (c) information technology, including advanced computing and communications; and (d) cognitive science, including cognitive neuroscience. [14, p. 1]

Yet the nano-ICT interface requires that some not insignificant problems be solved along the way. According to Kosta and Bowman [15, p. 257], human implant technology is likely to benefit from 'gradually more sophisticated nanofabrication techniques'.[4] These *nanophase* materials are, they note, pitched as offering 'a superior alternative to conventional orthopaedic implant materials', for example. The bio-nano interface raises important issues and problems. Some of these arise from developments in the use of nanomaterials, such as carbon nanotubes, for the neural interface in the field of bio-medical engineering. Research by He *et al* [16, p. 1], for example, investigated ways to engineer an electric interface between neural tissue and electrodes. This, they say, 'plays a significant role in the development of implanted devices for continuous monitoring and functional stimulation of central nervous system in terms of electroactivity, biocompatibility and long-term stability'. What health effects there may be as a result of these implants however remains uncertain.

When it comes to nano- or quantum computing, things are even more problematic, and there are many technical details to address in the first instance. As Tseng and Ellenbogen [17, pp. 1293-4] point out, the challenge of producing a 'commercially viable computer integrated on the molecular scale' requires that circuits are 'molecular scale in their entirety, not just incorporating molecular scale components'. This in itself raises a plethora of connected problems. For instance, it is true that the goal of assembling 'individual molecules of molecular-scale structures into functioning logic circuits' has already been achieved [17, pp. 1293-4], with 'electronic components including transistors, diodes, relays and logic gates from carbon nanotubes' [18, p. 5). Yet, the assembly of many more of these nanoscale structures engenders issues of, for example, excessive heat generation. As Kaewkamnerdpong and Bentley [18, p. 5] note, 'to build a molecular motor, researchers have to consider laws of thermadynamics when motors are actually in operation'. In addition to which there is the issue of how matter behaves at the nanoscale. 'The ability to manipulate individual atoms alone could not yet enable us to build reliable nanomachines, unless the physical principles at nanoscales are comprehended' [18, p. 3]. Resolving these issues will be important, not least because of issues of uncertainty and risk regarding the use of nanotechnology (more on this in Section 5 below)

Questions of how this technology might fruitfully be used is another question [17, pp. 1293-4]: 'The very small size of molecules make it possible, in principle, to fit a trillion molecular devices in a square centimetre. What does one do with a trillion devices?' How might this achievement be fruitfully utilised? One answer might be found in the field of swarm intelligence algorithms, employing evolutionary computational techniques that 'originated as a simulation of a simplified social system' [19, p. 81]. These systems, based on the group intelligence displayed by certain animals and insects such as birds, fish or ants, have proved effective in a number of programming areas. The intelligence capacity of individual members is not important; instead, it is the collaborative effort that is crucial. The potential of this programming approach for implants becomes most apparent when we consider the limitations of individual nanomachines by virtue of their very small size. One example of this in practice is the *Perceptive Particle Swarm Optimisation* (PPSO) algorithm, designed by Kaewkamnerdpong and Bentley [18, p. 9]:

[3] The ISO defines 'nanoscale' as 1-100nm, but I am provisionally using the term more generally to mean 1 nm to just under 1 micron (1,000nm).

[4] Information taken from Royal Society and Royal Academy of Engineering. 2004. *Nanoscience and Nanotechnologies: Opportunities and Uncertainties*. London: RS-RAE.

> Because each particle in the PPSO algorithm is highly simplified (each able to detect, influence or impact local neighbours in limited ways) and the algorithm is designed for working with a large number of particles, this algorithm would be truly suitable for programming or controlling the agents of nanotechnology (whether nanorobots, nanocomputers or DNA computers), whose abilities are limited, to perform effectively their tasks as envisioned.

The implications of these developments for implant technology cannot be underestimated. In the last decade the structure of implants has begun to change quite profoundly, and as a result of this convergence. Yet while there already exists interdisciplinary work between the fields (computer simulations in nano-research, development of nano-computer processors), and this is increasing, much of the focus seems to be on the development of research tools. For example, genetic algorithms, which 'can enable systems with desirable emergent properties'. They do this by promoting and selecting preferential methods and approaches and may prove useful in 'self-repair', for example. These sorts of 'genetic algorithms have [already] been used as a method in automatic system design for molecular nanotechnology' [18, p. 2].

More fundamental questions about what form future overlaps may take, and what effect they may have (whether within the independent fields, or more broadly on society, identity and health) have not however been considered in sufficient detail. What limited research there is can be found primarily within the sciences [18][20]. Nevertheless, funding for developing nanotechnology remains high on the agenda for many countries across the world, and the impact of this is likely to be far reaching. There are clearly issues to be raised about the manner and speed of such developments. For instance, questions about the possible toxicity of various nanomaterials remains unanswered, cf. [21], despite which, nanomaterials are already being employed within both health and industry (including, for example in household health and beauty products such as sun lotion). The implications of increasing overlaps between ICT and nanosciences requires sustained consideration, and on a number of platforms, not least in terms of legislation.

5 RISKS, BENEFITS AND LEGISLATION

As this paper has shown Nanomaterials, nanoelectronics, nano-computing, tissue engineering, as well as nano-enabled production processes, are converging in ways that raise philosophical, ethical, social and regulatory issues for which we are simply not prepared. Opinion tends to fall between those who believe it is likely to simply exacerbate existent risks, and those who think changes will be fundamental. Kosta and Bowman, for instance, [15, p. 257] predict changes are likely to be 'evolutionary, not revolutionary', thereby *amplifying* existent risks, whereas Fiedeler and Krings [11, p. 5] claim that the 'further penetration of technology into societal and cultural processes and vice versa' ought to be considered 'a deep transformation process'.

Either way, it is clear that when it comes to nanotechnology we are dealing with some very important uncertainties, and that more than one *kind* of uncertainty is involved [22]. This includes a lack of data, as might be expected considering its newness, but also more intrinsic uncertainties that result as a consequence of the complexity of living systems in their responses to nanoscale entities [22]. The trick is to account for this uncertainty in advance, and offer adequate legislation for future possible *unknown,* even potentially *unpredictable* risks?

The issue of long nanofibres (such as asbestos) is one such example of the risk inherent when manufacturing material with uncertain properties, specifically once they come into contact with bio-organisms. The challenge which Kosta and Bowman [15, p. 270] cite is to be able to regulate these potential risks (ethical, social, legal, human, environmental) 'against the broader public interest, while not compromising the development of a promising and powerful technology such as nanotechnologies'. The EGE report states that at the time of publication (2005) there were still 'no reliable scientific investigations concerning the long-term health impact of ICT implants in the body' [3, p. 23]. This gap persists; the technology is still very new. The authors therefore recommend 'proportionality of the tools that are used' in relation to the *purpose* of using implant technology [3, p. 19]. Such that 'even if the purpose as such is legitimate, it may not be pursued by using disproportionate tools'. Their recommendations for principles that ought to govern ICT devices for health purposes are [3, p. 30] that:

> a) the objective is important, like saving lives, restoring health or improving the quality of life;
> b) the implant is necessary to achieve this objective; and,
> c) there is no other less invasive and more cost-effective method of achieving the objective.

The problem with these objectives, however, is that they do not go far enough toward delineating boundaries between *improvement* and enhancement. This is even with a caveat that (a) would be with an aim to, say, 'improve the quality of life of people with severe injuries or conditions', since this presupposes a single measure regarding 'quality' of life, on which all people would naturally agree. Such agreement seems unlikely. Furthermore, the question of *cost-effectiveness* presupposes that ICT products will *remain* a costly luxury. With the introduction of lighter more economical nanomaterials becoming more of a possibility for ICT implants within the short term, this looks set to change. It is not that these issues are new, nor that they do not also arise in other areas of, for example, medicine or technology. The point is rather that the scale of uncertainty is a matter for concern.

For these reasons the authors of the report state [3, p. 2]: 'In its Opinion, the EGE makes the general point that non-medical applications of ICT implants are a potential threat to human dignity and democratic society'. In this account, *dignity* 'is used both to convey the need for absolutely respecting an individual's autonomy and rights and to support the claim to controlling individuals and their behaviour for the sake of values that someone plans to impose on other individuals' [3, p. 16]. They further note that *human dignity* 'concerns the self as an embodied self' [3, p. 28]. As such, this is an area requiring regulation, they claim, since, 'non-medical ICT implants in the human body are not explicitly covered by existing legislation' [3, p. 2], despite which, ICT implants may, in the future, lead to

the transformation of the human race' [3, p.28]. Perceptions of the human body as *data*, as opposed to complex social, cultural and natural beings, with the *potential for transformation*, has, they claim, 'large cultural effects' [3, p. 27]:

> particularly as it precludes higher level phenomena such as human psyche and human language or conceives them mainly under the perspective of its digitization, giving rise to reductionism that oversimplifies the complex relations between the human body, language and imagination.

In the light of this, we might conclude that a response balanced between cautious optimism and precaution is required when weighing up risks, hazards, costs and benefits of new technologies. Particularly when the potential of these converging technologies is yet to be even partially realised.

In the field of bio-medical engineering, nanomaterials such as carbon nanotubes can be used for the bio-nano neural interface. Yet the possible health implications of such implants are uncertain. One may raise questions about possible damage to DNA (even inter-generational), the immune and hormone system, protein-folding and generally about biocompatibility and bioaccumulation. The potential for so-called 'nanomachines', and the possible control of such machines using swarm intelligence (as already noted above) raises further questions (although, as noted above, these are not developments currently on the horizon). Nevertheless the speed of some developments means that we may not have time to develop strategies, nor gather knowledge about associated risks and assess potential harms in relation to benefits.

Those who would promote the benefits of technological advances might accept the need to minimise risks (ethical, social, legal, human, environmental) while being wary of 'compromising the development of a promising and powerful technology such as nanotechnologies' [15, p. 270]. A further difficulty in achieving this delicate balance is in answering the question: What constitutes risk, hazard or harm? As Wickson *et al* [23, p. 7] explain, 'While everyone may agree that scientists, policy makers and citizens should work to ensure that nanotechnology does not harm "nature" or "the environment", there are very different ideas about what these concepts mean, what constitutes harm, and the reasons why we might wish to avoid it'. Optimists like Kosta and Bowman [15, p. 271] believe that existing legislation, along with 'engineering based solutions' should suffice so long as researchers and manufacturers can be encouraged to consider integrating precautionary systems within their designs during the early stages of the products' development. These may include 'privacy-enhancing-technology or privacy by design within the nano-enabled ICT human implant system prior to its widespread employment' [15, p. 271]. This sort of optimism is not uncommon. Indeed, there are many other voices that question whether developers need to think of these issues at all. The belief is that legislation and ethical consideration should come later, and be left to others to think about. The difficulty with such claims is that, as McDonough and Braungar [24, p. 26] astutely reflect, 'At its deepest foundation, the industrial infrastructure we have today is linear: it is focused on making a product and getting it to a consumer quickly and cheaply without considering much else'. We simply cannot ignore the commercial aspect that drives some elements of research, nor ignore the fact that 'over time some technological practices become so entrenched in society that it becomes difficult to do things differently' [25, p. 213]. These factors combined mean we ought to consider such questions from the earliest stages of conception and design onwards, alongside related issues pertaining to freedom, dignity, privacy, security, consent, trust, control and equity of access.

As Feng [25, p. 213] notes, 'early on in the design process technologies are often malleable enough to be produced and implemented in a number of ways. Hence the need for ethical discussion to take place early on in the design of technologies'. Yet, and despite the abundance of questions offered above, current research on implants typically addresses either very general philosophical/ethical issues or practical issues (e.g. technical or regulatory) arising from these technologies. While the former is often discussed within philosophy, humanities and social sciences, the latter often comes from within industry, applied research, regulation and insurance. In fact, the absence of sufficient (and multi-disciplinary) discussion on these topics means that fundamental ethical and conceptual issues arising are not always considered, and thereby play insubstantial roles in enhancing the innovation and development of such technologies, both positively and by indicating limits.

How do we balance benefit and risk with regard to advancements in implant technology, and how far can, or should, existing regulation go (whether at pre- or post-production stage) with regard to the use of converging technologies for the development and use of implants? What, if any, role might *the precautionary principle* play here, and how might we regulate for future possible *unknown* risks without stifling technological and scientific creativity? I suggest that to fully engage with these issues we must proceed with both optimism and caution, and accept that thinking deeply, broadly and carefully about such matters means there may be no easy or quick answers.

ACKNOWLEDGEMENTS

Thanks are due to Prof. Geoffrey Hunt for his support in writing this paper, and to the reviewers whose feedback was enormously helpful. An earlier version of some sections of this paper were originally published in BioCentre's E-newsletter November 2011 edition. They are reproduced here with kind permission. A copy of that paper can be found here: http://www.bioethics.ac.uk/news/ICT-Implants-nanotechnology-and-some-reasons-for-caution.php

REFERENCES

[1] Plato, *Phaedo*, any edition.

[2] Decartes, R. (1968). *Discourse on Method and the Meditation*, Middlesex: Penguin. (6th Meditation)

[3] European Group on Ethics (EGE) 'Opinion No 20: Ethical Aspects of ICT Implants in the Human Body'. Presented to the European Commission in March 2005.

[4] Clark, A., Chalmers, D. J. (1998) The Extended Mind *Analysis*. 58. 10-23.

[5] De Preester, H. (2011). Technology and the Body: the (Im)Possibilities of Re-embodiment. *Foundations of Science*. 16. 119-137.

[6] Nsanze, F. (2005). ICT implants in the human body: a review. In *Opinion No 20: Ethical Aspects of ICT Implants in the Human Body*. Presented to the European Commission in March 2005.

[7] Lagali P., Corcoran C., Picketts, D. (2012). Hippocampus development and function: role of epigenetic factors and implications for cognitive disease. *Clinical Genetics,* 78: 4. 321-333. (p.326).
[8] Graham-Rowe, D. (2003). 'World's first brain prosthesis revealed', *New Scientist* 12 March 2003. Online: http://www.newscientist.com/article/dn3488-worlds-first-brain-prosthesis-revealed.html [accessed 27/06/11].
[9] Soekadar, S., Haagen, K., Birbaumer, N. (2008). Brain-computer interfaces (BCI): restoration of movement and thought from neuroelectric and metabolic brain activity. In *Coordination: Neural, Behavioral and Social Dynamics*. 229-252.
[10] Carmena, J. M., Lebedev, M. A., Crist, R. E., O'Doherty, J. E., Santucci, D. M., Dimitrov, D. F., Patil, P. G., Henriquez, C. S., Nicolelis, M. A. L. (2003). Learning to control a brain-machine interface for reaching and grasping by primates. *PLoS Biology*. 1: 193-208.
[11] Fiedeler, U. and Krings, B. J. (2006). Naturalness and neuronal implants – changes in the perception of human beings. MPRA Paper presented at *EASST-conference*, University Library of Munich, Germany.
[12] Burger, R. (2002). Enhancing personal area sensory and social communication through converging technologies, in Roco, M. C. and Bainbridge, W.S. (eds.).
[13] Strydis, C., Gaydadjiev, G. N. (2008). The case for a Generic Implant Processor. *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'08),* August 2008. 3186–3191.
[14] Roco, M. C., Bainbridge, W. S. (eds.). (2002). Converging technologies for improving human performance: nanotechnology, biotechnology, information technology and cognitive science. *National Science Foundation,* Arlington, Virginia, USA. http://www.wtec.org/ConvergingTechnologies [accessed 01/07/11].
[15] Kosta, E., Bowman, D. M. (2011). Treating or tracking? regulatory challenges of nano‐enabled ICT implants. *Law and Policy.* 33: 2. 256-275.
[16] He, L., Lin, D., Wang, Y., Xiao, Y., Che, J. (2011). Electroactive SWNT/PEGDA hybrid hydrogel coating for bio-electrode interface, *Colloids and Surfaces B: Biointerfaces.* 87: 2. 273-279.
[17] Tseng, G. Y., Ellenbogen, J. C. (2001). Toward nanocomputers. *Science*. 294: 5545. 1293-1294.
[18] Kaewkamnerdpong, B. and Bentley, P. J. (2005). Computer science for nanotechnology: needs and opportunities. *Proceedings of the 5th International Conference on Intelligent Processing and Manufacturing of Materials (IPMM)*. Online: http://www.cs.ucl.ac.uk/staff/b.kaewkamnerdpong/bkpb-ipmm05.pdf [accessed 20/06/11].
[19] Eberhart, R. C. and Shi, Y., (2001). Particle swarm optimisation; developments, applications and resources', *Proc. IEEE Int. Conf. Evolutionary Computation*. 1: 81-86.
[20] Heath, J. R., Kuekes, P. J., Snider, G. S., Williams, R. S. (1998). A defect-tolerant computer architecture: opportunities for nanotechnology. *Science*. 280: 5370. 1716-1721.
[21] Howard, C. V. and Ikah, S. K. (2006). Nanotechnology and nanoparticle toxicity: a case for precaution. In Hunt, G. and Mehta, M. *Nanotechnology: Risk, Ethics and Law*, 2006. London: Earthscan. 154-166.
[22] Hunt, G. and Riediker, M. (2011). Building expert consensus on problems of uncertainty and complexity in nanomaterial safety. *Nanotechnology Perceptions.* 7: 2. 82-98.
[23] Wickson, F., Grieger, K., Baun, A. (2010). Nature and nanotechnology: science, ideology and policy. *International Journal of Emerging Technologies and Society*. 8: 1. 5-23.
[24] McDonough, M. and Braungar, M. (2002). *Cradle to cradle: remaking the way we make things*. New York: North Point Press.
[25] M. Feng, P. (2000). Rethinking technology, revitalizing ethics: overcoming barriers to ethical design. *Science and Engineering Ethics*. 6: 2. 207-220.

Cantor's Diagonalization and Turing's Cardinality Paradox

Dale Jacquette[1]

Abstract. *Turing's Cardinality Paradox* is the result of two opposed but comparably powerful lines of argument, supporting the diametrically opposed propositions that the cardinality of dedicated Turing machines outputting all and only the computable binary digital sequences can only be denumerable, and yet must also be nondenumerable. Turing's criticisms of other applications of diagonalization are addressed, and directions for further research in avoiding the paradox and understanding the concept of a Turing machine, computability and computable real number sequences are considered.

1 COMPUTING MACHINE CARDINALITY

A.M. Turing's [1] concept of computability, computing machines, and computable binary digital sequences, is subject to Turing's Cardinality Paradox. The paradox results from two opposed but comparably powerful lines of argument, supporting the propositions that the cardinality of dedicated Turing machines outputting all and only the computable binary digital sequences can only be denumerable and yet must also be nondenumerable. Turing's objections to a similar kind of diagonalization are answered, and the implications of the paradox for the concept of a Turing machine, computability, computable sequences, and Turing's effort to prove the unsolvability of the *Entscheidungsproblem* are explained in light of the paradox.

Circle-free Dedicated Turing Machines DTMs symbol-edit their way algorithmically, over ideal infinite time, to produce all and only the computable sequences, CSs, correlated with all and only computable real numbers. They do so, even if, as automatic machines, in Turing's sense, they cannot be made to do anything more. Turing pronounces on the cardinality of the CSs and by implication on the cardinality of the circle-free DTMs in these terms: 'To each computable sequence there corresponds at least one [machine] description number, while to no description number does there correspond more than one computable sequence. The computable sequences and numbers are therefore enumerable' [1, p. 241].

Turing backs up this conclusion concerning the denumerable cardinality of the CSs with an intriguing independent argument. He holds on preceding pages leading up to his assertion above that there is a unique positive integer coding, a standard machine description number (S.D.), for every DTM, circle-free and circular alike, which he explains as being constructable from each DTM machine instruction table, in somewhat the way that Gödel numbers by glossary of basic terms code any construction in the syntax of a symbolic logic. The machine instruction table of each DTM is itself a finite symbol string, capable of being integer coded as a rather large number, and can accordingly be printed as a CS by a DTM. There are only denumerably infinitely many integers. If the circle-free DTMs are a proper subset of the totality of DTMs, and if there can only be as many DTMs altogether as there are positive integers by which each DTM is coded, then there can only be denumerably infinitely many DTMs. Hence, Turing maintains that there can be at most denumerably infinitely many circle-free DTMs, from which it further follows that there can equally be only denumerably infinitely many CSs [1, pp. 239-241].

[1] University of Bern, Switzerland, email: dale.jacquette@philo.unibe.ch

2 TURING'S EXPOSURE OF A DIAGONALIZATION 'FALLACY'

It is possible nevertheless to demonstrate, contrary to the above denumerability requirement, that the cardinality of all and only circle-free DTMs and all and only CSs is actually *non*denumerable.

Turing proposes to expose this kind of diagonalization as depending on a 'fallacy'. It is argued in what follows that, even if Turing is right that the diagonalization he considers is fallacious in just the way he says, there is another version of the same style of diagonalization that does not commit the fallacy, and that, properly interpreted, implies that in Turing's world the circle-free DTMs and CSs can only be nondenumerably infinite in cardinality. Turing offers a concise statement of the diagonalization principle in criticizing one especially frontal type of diagonalization on the surjectively (many-one) correlated DTMs and CSs, respectively. In [1, §8], Turing anticipates the following version of a Cantor-inspired diagonalization applied to supposedly complete denumerably infinite listings of all and only the CSs, and by extension concerning all and only the circle-free DTMs:

> It may be thought that arguments which prove that the real numbers are not enumerable would also prove that the computable numbers and sequences cannot be enumerable [...] Or we might apply the diagonal process. If the computable sequences are enumerable, let α_n be the n-th computable sequence, and let $\phi_n(m)$ be the m-th figure in α_n. Let β be the sequence with $1-\phi_n(n)$ as its n-th figure. Since β is computable, there exists a number K such that $1-\phi_n(n) = \phi_K(n)$ [for?] all n. Putting $n = K$, we have $1 = 2\phi_K(K)$, i.e. 1 is even. This is impossible. The computable sequences are therefore not enumerable. [1, p. 246]

If Turing's remarks apply to the diagonalization proposed, then so should his solution. Turing argues that the diagonalization fails because it embodies the false assumption that diagonal β is computable, as he continues:

> The fallacy in this argument lies in the assumption that β is computable. It would be true if we could enumerate the com-

putable sequences by finite means, but the problem of enumerating computable sequences is equivalent to the problem of finding out whether a given number is the D.N. [machine or m-description number] of a circle-free machine, and we have no general process for doing this in a finite number of steps. In fact, by applying the diagonal process argument correctly, we can show that there cannot be any such general process. [1, p. 246]

From the essay, it is hard to judge whether Turing means to deflect diagonalization as a friend or foe of DTMs. The context and Turing's word choice suggest that he would be happy to avail himself of the Cantorian diagonalization argument he mentions in order to support the same conclusions about the *Entscheidungsproblem* for distinguishing circular from circle-free, if it only it were not subject to the nuisance fallacy he identifies [1, p. 246].

Turing afterward offers a different diagonalization that avoids the objection, but has a different purpose, from which Turing in a series of increasingly interesting inferences deduces the mechanical unsolvability of the *Entscheidungsproblem*. It appears, on the contrary, that if a Cantor-style diagonalization on the CSs and surjectively correlated DTMs of the type to be described succeeds, then Turing's concept of computability, of a CS and DTM, is landed squarely in inconsistent cardinality attributions to the domain of all CSs and all circle-free DTMs, to which Turing must simultaneously ascribe both denumerably and nondenumerably infinite cardinality.

Turing's objection presupposes that in projecting a diagonalized CS, β, relative to a listing of binary digital sequences, CSL_α (simply α, in Turing's original symbolism), we must have already solved the *Entscheidungsproblem*, so as to know whether or not a given DTM is circle-free by having determined that its standard machine description number is satisfactory. If Turing succeeds in showing that the *Entscheidungsproblem* is algorithmically unsolvable, then that fact would evidently be a weighty impediment to the exact use of diagonalization to which Turing objects. Turing's proof of the unsolvability of the *Entscheidungsproblem*, unfortunately, depends on the assumption that the CSs and DTMs are denumerable, and that is precisely the assumption challenged by a Cantor-inspired algorithmic diagonalization on the CSs in any CSL. Furthermore, the diagonalization target for the argument is a listing CSL_α, supposedly, of all and only the denumerably infinite CSs computed by all and only the denumerably infinite circle-free DTMs. Turing at first appears right to maintain that we cannot posit or project a listing of all circle-free DTMs or the integers (themselves also real numbers) that code the standard machine descriptions of all circle-free DTMs, without having already solved the *Entscheidungsproblem*, because we cannot know in advance of solving the decision problem which sequences are computable by DTMs, and which are not. Thus, we cannot project a listing of all and only the circle-free DTMs, rendering β uncomputable, and, as such, not after all a CS. To proceed in any such fashion and conclude from these assumptions that there are nondenumerably many CSs or DTMs would certainly be fallacious, as Turing maintains.

The fallacy is nevertheless avoidable, and does not affect every Cantor-style diagonalization that might be offered on listings of CSs. All that need be argued is not that we can solve the *Entscheidungsproblem*, or even that the *Entscheidungsproblem* is solvable. To prove that there are nondenumerably infinitely many circle-free DTMs and corresponding output CSs, we need only show that there exists at least one diagonal CS or DCS that cannot belong to any *reductio*-hypothetically complete and denumerably infinite listing of all and only the CSs output by all and only the denumerably infinitely many circle-free DTMs.

3 TURING'S CARDINALITY PARADOX: DIAGONALIZATION OF COMPUTABLE SEQUENCES

Turing holds that if the sequences belonging to α are computable, then a computable diagonal operation on the sequences in α is also possible, and in this he is certainly right. What does not follow is the inverse conditional that Turing by his reasoning needs in order to block the diagonalization, that if the sequences belonging to α are *not* computable, then a diagonal operation on the sequences is *not* possible. The denumerable circle-free DTMs and CSs might be surjectively correlated in any supposedly complete listing, here α:

$$\begin{aligned} \mathrm{DTM}_1 &- 0011001100110011\ldots \\ \mathrm{DTM}_2 &- 0101010101010101\ldots \\ \mathrm{DTM}_3 &- 1101101101101101\ldots \\ &\vdots \end{aligned}$$

We define a diagonal DTM or DDTM on any CSL_α, taking as its <basis> any DTM_k occurring in any row $k \geq 1$ in CSL_α, CSL_α[DDTM<DTM_k>], and outputting a diagonal CS or DCS relative to CSL_α. A DDTM is a DTM that computes an interpretation of another subordinate DTM; or, more precisely, scans and implements the machine instructions of the interpreted DTM, so as algorithmically to modify the CS_1 of the interpreted DTM in a particular way, resulting in another CS_2 than the subordinate DTM would have computed on its own. What follows, accordingly, is the most direct and obvious application of Cantors diagonalization to the CSs, and, by implication, to the DTMs. The algorithm for DDTM, adopting some of Turing's 'abbreviated' or 'skeletal' machine instruction shorthand, the effect of DDTM<DTM_k> in CSL_α[DDTM<DTM_k>], is computed according to the following traditional Cantor-inspired diagonalization style, reading $MI(d_k) = P(0) \rightarrow P(1)$ as saying, conditionally, that if DTM_k is machine instructed to print 0 at digit place k in the CS it computes, then DTM DDTM<DTM_k> is to print 1 (instead):

Diagonal D-Rule CSL_α[DDTM<DTM_k>] = $\mathrm{DTM}_k[P(d_1), \ldots, P(d_{k-1}); MI(d_k) = P(0) \rightarrow P(1); MI(d_k) = P(1) \rightarrow P(0); P(d_{k+1}), P(d_{k+2}), \ldots]$

If, as we expect, we can mechanically apply DTMN<DTM_1> to output an algorithmic modification to the CS in any row k of any listing CSL, even without algorithmically identifying digit place k in CS(DTM_1), then DTMN is a diagonal DTM, a DDTM, for any circle-free DTM outputting any CS in any row of CSL.

Turing cannot accept that the CSs and circle-free DTMs are of transfinite cardinality, because he adopts what is arguably the inadequately supported, and on its own certainly logically open, conclusion, that all DTMs can be integer coded. It might be preferable to say, especially if Turing demonstrates the unsolvability of the *Entscheidungsproblem* applied to standard machine description numbers, if Turing is right in believing he has demonstrated that there is no algorithmic singling out of all and only the satisfactory numbers coding all and only the circle-free DTMs, that we cannot ascertain even and especially as a matter of mathematical principle that the

circle-free DTMs can be denumerably integer coded, unless or until we have independently ascertained that the circle-free DTMs are denumerably infinite. Otherwise, Turing's integer coding argument for the denumerability of the circle-free DTMs blatantly begs the interesting question.

We can variously express these diagonalization results involving DDTMs, relating them to the cardinality $\mathfrak{C}$ of the totality of DTMs and of the CSs, or to the denumerably infinite totality of natural numbers, $\mathbb{N}$ (or whole numbers, integers or positive integers), for both the set of all dedicated Turing machines, DTMs, and the set of all CSs. Turing's Cardinality Paradox can then be succinctly expressed as the inconsistency by which the DTMs are at once collectively in their totality denumerably and nondenumerably infinite in cardinality: $\mathfrak{C}(\mathrm{DTMs}) = \aleph_0 \wedge \mathfrak{C}(\mathrm{DTMs}) > \aleph_0$; $\mathfrak{C}(\mathrm{DTMs}) = \aleph_0 \wedge \mathfrak{C}(\mathrm{DTMs}) \neq \aleph_0$ (and similarly for $\mathfrak{C}(\mathrm{CSs})$ by virtue of the surjective correlation of CSs and DTMs generally and circle-free DTMs in particular).

ACKNOWLEDGEMENTS

I am grateful to my research assistant Jan Walker for invaluable discussions of Turing's classic paper and surrounding literature. The argument has benefited substantially from Jan's energetic criticisms of previous versions.

REFERENCES

[1] A.M. Turing, 'On computable numbers, with an application to the entscheidungsproblem', *Proceedings of the London Mathematical Society*, **42**, 230–265, (1936).

The Proof Theoretic Foundations of Computation with Application to Turing's Thesis and the Chinese Room Argument

Michael Gabbay [1]

Abstract. This paper aims to provide a proof theoretic account of computation, and so extend the proof theoretic programme of accounting for logical constants and laws to account for computational constants and laws. The notion of a computational law can then provide an abstract characterisation of a computer and a program.

The formal theory of computation obtained (proof theoretically) is then applied to two famous questions in the philosophy of computation: a new argument for Turing's Thesis; and a new defence of computationalism against the Chinese Room argument.

Sections 1 and 2 introduce and motivate the proof theoretic account of logic. Section 3 discusses the concept of computation, and Section 4 develops a proof theoretic analysis of it. Section 5.1 uses this analysis to argue for Turing's Thesis, and Section 5.2 offers a further application to a defence of computationalist theories of the mind against Searle's famous Chinese room argument.

1 Introducing Proof Theory

There are many versions of the proof theoretic account of logic. Or at least, elements and variations of it appear in much philosophical literature on the nature of logic. The origin of the proof theoretic treatment of the logical constants is commonly cited to be Gentzen's remarks in [10], where he suggests that certain inference rules may be seen as definitional of their subject connectives.

Gentzen's ideas were famously taken up by Prawitz, Hacking and Dummett (among others, e.g. Brandom). They argued that being defined by its derivation rules is what it is for a connective *to be logical*. Hence Hacking's answer to the question "what distinguishes logic from the extralogical?" is a proof theoretic one: the logical connectives are exactly the ones which obtain their meanings from their derivation rules.

But Gentzen was not the first to offer an account of logic founded on proof-theory. Frege, at the beginning of [5], defines a truth of logic as one which can be *derived* from some 'primitive' truths. Moreover, Frege was not concerned only with logic. The theory developed in [5] and [6] is primarily a foundational theory of *functions*. Frege's goal — a goal that is to an extent maintained by set theory and category theory — was to treat the extensions of mathematical and logical terms as particular functions within his general framework. For example, the extension (i.e. 'reference' or 'denotation') of a sentence, according to Frege, is a truth value (a particular constant function) and the extension of a sentential connective is a function on truth values; so the extension of $A \wedge B$ is the extension of $\wedge$ *applied to* the extensions of A and B (just as we commonly think of the extension, or reference, of 3+2 as being what results from applying the extension of + to the extensions of 3 and 2).

Because of the logical paradoxes, in particular Russell's paradox, Frege's general theory of functions was largely abandoned. With the exception of some ideas of Church, which encountered similar paradoxes to Frege's theory (see, e.g. [14]), logic and computation have been regarded as separate areas with distinct, although connected, foundations.

This paper offers a proof theoretic foundation for a theory of functions by appealing to an extension of classical sequent calculus that can reason on computational 'constants'. Coupled with a proof theoretic account of logic this paves the way for a single foundation for logic and computation.

The sense in which this paper offers a 'single foundation' for logic and computation is somewhat narrower than the sense in which Frege can be said to have offered one. This paper shall be concerned with providing an account of some general laws governing functions that are sufficient to describe computability. We do not posit an intricate universe of functions into which the entirety of mathematics can be reduced. Furthermore, the foundation of logic and computation is 'single' in the sense that the same strategy, a proof theoretic one, is used to account for logical and functional laws. It is not an attempt to reduce logic to a theory of functions or vice versa.

2 Proof theory and logic

As noted in Section 1 there is no one proof theoretic treatment of the logical constants. Here we work with a version of it, taken from Hacking's paper [11]. Henceforth this paper will refer to 'the proof theoretic account' meaning Hacking's version of it.

The proof theoretic account (of classical logic) begins, not with proof theory, but with some semantic principles.

> The subject matter of logic is consequence relations; logical consequence is a relation on propositions; propositions are either true or false; and consequence relations preserve truth. (1)

[1] Department of Philosophy Kings College London

These principles are then characterised proof theoretically, by the following structural rules:[2]

$$\frac{}{A \vdash A} \quad \text{(Reflexivity)}$$

$$\frac{\Gamma \vdash \Delta}{\Gamma, A \vdash \Delta} \qquad \frac{\Gamma \vdash \Delta}{\Gamma \vdash A, \Delta} \quad \text{(Weakening)}$$

$$\frac{\Gamma \vdash A, \Delta \quad \Gamma', A \vdash \Delta'}{\Gamma, \Gamma' \vdash \Delta, \Delta'} \quad \text{(Cut)} \tag{2}$$

Γ, Δ are *sets* of propositions

The connection between (2) and (1) is closer than one might initially think. For example, (Cut) and (Reflexivity) are clearly connected to the reflexive-transitive relation of truth preservation. Also, as has been shown in [1], there is a close connection between the remaining structural rules and the condition that there are only two truth values. So there is already good formal support for the claim that (2) is a proof theoretic characterisation of (1).

The proof theoretic account continues by specifying that a logical connective is one that is definable, by its inference rules, so as to preserve the rules of (2) (e.g. [11, p.296]).

To see what 'preserve' means here we must think of the rules of (2) as being laws in a Humean sense: merely describing regularities. We could have a formal system in which these regularities hold, but then lose these regularities when connectives with new inference rules are added.

A famous example is Prior's connective tonk, which is 'defined' by the inference rules:

$$\frac{\Gamma, A \vdash \Delta}{\Gamma, A \text{ tonk } B \vdash \Delta} \ (\text{tonk}L) \qquad \frac{\Gamma \vdash B, \Delta}{\Gamma \vdash A \text{ tonk } B, \Delta} \ (\text{tonk}R)$$

Imagine a tonkless formal system where it just so happens that when$\Gamma \vdash A, \Delta$ and $\Gamma', A \vdash \Delta'$ are derivable, then so is $\Gamma, \Gamma' \vdash \Delta, \Delta'$ (possibly by a completely different derivation). Now, except in trivial cases, the addition of the rules for tonk do not preserve (Cut). This is because we can use them to derive the premises of an instance of (Cut) without being able to derive its conclusion.

The matter is confused somewhat by the standard reference to (2) as describing structural 'rules'. The account makes more sense when (2) is regarded as describing (accidental) *regularities*, rather than rules. Thus, according to the proof theoretic account, inference rules for some logical connectives must be such that they happen to satisfy the *regularities* of (2).

The requirement that logical rules should preserve (2) is now justified in terms of the semantic principle (1) to which (2) corresponds. Assuming (1) as a general semantic principle, anything that has a meaning at all must be compatible with (1). One test of compatibility with (1) in the case of inference rules is that they preserve the structural regularities to which (1) corresponds: i.e. (2). The proof theoretic account then defines a system of connectives as *logical* when all their inference rules preserve (2).

Notice that this does not entail that a term is meaningless if its inference rules do not preserve (2), or if it has no particular inference rules. Such terms, of which there are many in all natural languages, are simply not logical.

To repeat, on the proof theoretic approach, the semantic principle (1) justifies the structural rules (2) as long as we know that the symbols involved are defined (or meaningful). A connective is *logical* when it is defined by its inference rules. Inference rules can define a connective when they preserve the admissibility of the rules of (2) (in any system where (2) is satisfied by sentences not containing the defined connective). Finally, logic is the theory of what is derivable or definable using logical connectives.[3]

The proof theoretic account presented here only goes so far as to account for which terms are logical and why. It does this given some basic semantic principles which, in a sense, already determine the logic. But it is by no means trivial what this sense is, it is the rather subtle proof theoretic sense developed by the proof theoretic account.

Some authors, e.g. [2] and [15], have taken the proof theoretic account a step further. They offer a non-semantic justification of rules like (2), e.g. by appealing to primitive inferential practices, and then use that to justify (1). Such accounts attempt to give a completely proof theoretic account not only of which terms are logical, but also of which logic (classical, intuitionist etc.) is *the* correct logic. This paper does not go so far as these authors, and does not address objections to (1) as being unjustified or untenable. The point of the paper is to show that his approach can be extended from logic to computation.

It is worth noting that the proof theoretic approach yields much from few resources. From the basic semantic principles of (1) we obtain the whole of first order quantified logic, without ever mentioning models, valuations, properties etc.. Exactly how much the proof theoretic account yields depends on what one chooses to count as a derivation (or proof). For example, Hacking saw only infinitary proof methods as a means of obtaining an inferential definition of the terms of arithmetic. Dismissing infinite inference rules as not being genuine definitions, Hacking concluded that:

> Recursive arithmetic cannot be reduced to logic, strictly construed. [11, p.316]

Given Turing's Thesis, which equates computable and recursive functions, Hacking would probably have concluded that computability cannot be given a proof theoretic, logical, justification. In what follows it shall turn out, among other things, that Hacking's conclusion was too hasty.

3 What is a computer?

3.1 Computers and computable functions

We now turn to the question: 'what is computation and what is a computer?'. These questions are hard to answer separately as the two concepts are deeply intertwined: a computer is something that carries out computations and a computation is something that a computer does.

The strategy I propose for dealing with this difficulty is to provide an abstract definition of a *computable function*. A

[2] The symbol $\vdash$ represents the relation of logical consequence. It's fundamental content is, according to the proof theoretic account, merely that of a truth preserving relation as specified in (1).

[3] Given some putative logical connectives, the hardest admissability result to obtain is generally the admissibility of (Cut) (commonly, and confusingly, called *cut-elimination*).

computer can then be regarded as any machine that determines the values of computable functions, and computations are the relevant steps in the machines' processes.

There are two further directions we could take now: we could follow Turing and provide a 'mechanical' definition of a computable function (in terms of e.g. Turing machines); or we could follow Church in providing a 'mathematical' (in terms of an abstract formal theory of functions).

But before continuing it is worth addressing a natural objection that, according to this strategy, too many things become computers. For example any object, e.g. a fingernail, becomes a computer that 'computes' whichever constant function we choose that object to represent.

To answer to this objection we should note that a machine determines the value of a function when it is built so that it correctly manipulates the symbols representing its domain and range of the function. Thus a computer on this view is not simply a machine, it is a machine together with an interpretation of the symbols it manipulates. In this sense, a fingernail can be a computer when associated with an appropriate representation of a constant function. Suppose we use a fingernail to represent the constant function 1. Then the fingernail itself is equivalent to a Turing machine that asks for no inputs and outputs only a symbol representing 1.

Notice that a machine M is a computer, on this account, when appropriate relations hold between whatever are interpreted as Ms numerical inputs and its outputs. If we can give an independent characterisation of what these functional relations are (i.e. what the computable functions are), then we are done.

3.2 The classical model of computation

Turing's approach to characterising computable functions is to consider a special kind of 'discrete-state machine': electromechanical devices with symbolic memories, processors and a command table; we can call these 'Turing machines'. Turing could then give very precise and concrete descriptions of the architecture and operation of Turing machines as symbol manipulation devices.

If we hypothesise that any computatable function can be determined by a Turing machine, then, following the proposed strategy, we can then define a computer to be any (physical) system that can be simulated by one of these machines. The hypothesis on which such a definition of 'computer' depends can be called the *Turing's Thesis*.

So machine M is a computer on this view when it can be emulated by a Turing machine (that manipulates symbols that represent Ms inputs and outputs). Exactly what goes on 'inside' M is not significant, all that is significant is that it determines the values of functions as a Turing machine could.

It is this sense of a computer that best motivates Turing's suggestion in [19] that

> the question, "Can machines think?" should be replaced by "Are there imaginable digital computers which would do well in the imitation game?"[4]

For if 'computer' means 'emulatable by a Turing machine' then a thinker that is emulated by a Turing machine is also a computer, and so is an example of a thinking machine (albeit an organic one).

The characterisation of a computer as being a mechanism that can manipulate symbols, as a Turing machine can manipulate them, is often referred to as the *classical* model of a computation. It is this conception of computers and computation that underlies many computational theories of the mind. For example, in a recent work criticising the computational theory of the mind, Fodor writes that

> in Classical models, the architectural processes ... all are (or reduce to) operations defined over symbols that belong to the primitive vocabulary of the language that the machine computes in (they are operations like, e.g., writing a primitive symbol, deleting a primitive symbol, and the like). [4, p.45]

Clearly, in Fodor's picture, a computer is something that manipulates symbols according some predetermined rules: a variant of a Turing machine.

3.3 Problems with the classical approach

I wish to highlight two difficulties with the characterisation of computers as being machines emulatable by Turing machines.

The first difficulty relates to the status of the fundamental hypothesis that all computable functions can be determined by a Turing machine, Turing's Thesis. Although there is significant empirical and mathematical evidence for it, it would be useful if our characterisation of computers and computation should shed light on its truth, rather than assume it. I shall return to Turing's Thesis in Section 5.1.

A second difficulty with this 'classical' approach is that it may provide an answer to the question 'what is a computer?', but does it does not answer the question 'what is *computation*?'. I will now discuss this problem in more detail.

Intuitively, a computer is something that 'computes' values. But a computation is the implementation of a program in a computer. We could have two computers that are equivalent in the sense that they halt with the same outputs for the same inputs (i.e. are Turing-equivalent, they emulate each other), but they run different programs. A variation on a famous argument in the philosophy of mind will help to make this point.

Given that we only ever encounter machines of finite size, and have finite time to run them, it is hypothetically possible to catalogue all the outputs from all possible inputs of a computer over a fixed period of time (and how long the computer takes to produce these outputs). Suppose that we have two rooms, in one is a commercial supercomputer, in the other is a person with a catalogue of the behaviour of that supercomputer. Both rooms are fed inputs, the supercomputer does what it does to them, and the person in the room looks up the inputs in his catalogue and as the catalogue instructs, he returns the corresponding outputs. The two rooms may be indistinguishable, and we may even wish to say that both rooms contain supercomputers, but the computations that are occurring within the two rooms are clearly different. One is engaging in billions of calculations a second, the other is running a look-up table.[5]

[4] As Turing notes in his paper, digital computers can emulate and be emulated by the Turing machines.

[5] For simple calculations the person with the table will be slower

The example just given is a variation on Searle's famous Chinese room argument of [16]. In Searle's argument we are asked to consider a Chinese speaker on the one hand and somebody mindlessly simulating a Chinese speaker, by means of a look-up table, on the other. Searle uses his argument to conclude that various forms of computationalism are false. In particular, the form proposed here: that mental states are particular programs running on a computer; and programs themselves are implementation independent (see e.g. [12]). If by this it is meant that Turing-equivalent computers run the same programs, then it follows that the real Chinese speaker and the room have the same mental states. But this is intuitively false.

Questions of whether mental states are computer programs aside, there is a strong intuition that whether two programs are the same is a hardware independent matter. What the two examples above have shown is that Turing-equivalence is not a good test for whether two computers are running the same programs.

There is an analogy here with logic. Consider the these two sets of propositions: $\{A \wedge B\}, \{\neg A \vee B, A\}$, these two sets are equivalent in the sense that each can be derived from the other. However, they do not contain the same propositions. Indeed, in some branches of computer science, e.g. *logic programming*, propositions are seen as types of program, and this analogy is very close.

This second difficulty arises in a slightly different form if we ask what a *computational step* is. An intuitive characterisation of computational steps is given in terms of the architecture of a discrete state computer such as a Turing machine. A computational step is one 'click' in the Turing machine — the machine writing a symbol and then moving the tape one cell left or right — and a rule of computation is an instruction of what steps the Turing machine should take depending on what is on its tape. This account is problematic in two respects. Firstly it is not implementation independent as it makes reference to the specific hardware of Turing machine. Secondly, and perhaps more seriously, not every discrete step in the action of a computer is a *computational* step. To see this, consider a Turing machine programmed to perform some complex operation on numbers. In the case of a Turing machine certain cells on the tape will be used to store symbols for later parts of the calculation. Then some of the operations of the Turing machine will simply be to move the 'head' away from 'stored' symbols so that no new calculations overwrite old ones (which will be needed later). Now, such operations are not computational steps, they are required for the program to run correctly, but are not actually parts of the computation.[6]

The point is that not every step in the process of a computer carrying out a computation is a *computational* step. We therefore seek a hardware independent, abstract characterisation of a computational step. Then we can define a program as something that puts into effect some particular sequences of computational steps, and define a computer as something that runs programs. It is hard to see how this could be done in sufficient generality in terms of Turing machines (or any other piece of hardware).

than the supercomputer. To avoid this matter we need only replace the person with some machine specially designed to look things up quickly.

[6] As another example, suppose that after every n moves of the tape, the Turing machine engages a fan system, or oils itself to insure smooth operation. These are essential parts of the machine running a computation, but are not themselves *computational*.

4 Proof theory and computation

I now turn to an alternative, logical, account of what a computable function is. I will propose a logic of computation, justifiable in a similar way to the justification of the logic of propositions described in Section 2. This account, I claim, can better handle the two difficulties above. In particular, we can use it to account for Turing-equivalent but distinct computations, and to give an argument for the truth of Turing's Thesis.

4.1 Proof theory and untyped λ-calculus

As observed already, the proof theoretic account of logic offers an account of what the logical inference rules are and why they are logical. It turns out that we can use proof theory similarly to account for computational laws.

As with logical consequence, a proof theoretic account of computation begins with semantic principles.

> The relata of computational relations are functions; functions take other functions as their values and arguments; and computing on a function preserves its values. (3)

To gain intuition for this principle, consider a simple computation on the function (2×3)+(1+0): 0,1,2, and 3 are (constant) functions; 2×3 is the function '× applied to 2 and 3'; 2×3 computes to 6; 1+0 computes 1; and so the whole thing computes to the function 7. As instantiated in this simple example, each computational step preserves the overall value of the function. In this case the function computes to a constant value, 7, but some computations are far more complex and may not terminate so easily (or at all).

The semantic principles of (3) seem well characterised proof theoretically by the structural rules of the term sequent system developed in [8],where a derivation system is developed for *untyped* λ-calculus. Let us use $\rightsquigarrow$ to represent to relation '... computes to...', then, still somewhat informally, the structural rules (or regularities) are corresponding to (3) are:

$$\frac{}{t \rightsquigarrow t} \quad (\lambda\text{-Reflexivity})$$

$$\frac{t \rightsquigarrow s \quad \text{and} \quad g(\dots s \dots) \rightsquigarrow r}{g(\dots t \dots) \rightsquigarrow r} \quad (\lambda\text{-Cut}) \tag{4}$$

The rules (λ-Reflexivity) and (λ-Cut) correspond to the principle that computation preserves value. For example, (λ-Cut) can be read informally as saying that if t features in a more complex function $g(\dots t \dots)$, and t computes to s, then computing t 'inside' $g(\dots t \dots)$ preserves the value of $g(\dots t \dots)$ (as r). The structure of the sequent itself, allowing function symbols to apply unrestrictedly to other function symbols, corresponds to the semantic principle that we are dealing with (total) functions that can take any other functions as arguments. The system is described in more detail in the appendix to this paper.

In much the same way that rules are given for logical quantifiers, we can present rules for an (untyped) function abstraction operator λ. The λ operator has the following intuitive

meaning: if the term t denotes the value of some function at value x, then the term $\lambda x.t$ represents that function f. So for example, $x{+}2$ represents the result of adding 2 to x and so $\lambda x.(x{+}2)$ represent the function of 'adding 2'. The most important rule of the λ-calculus is that of *β-reduction*:

$$\lambda x.t(s) \text{ computes to } t[x/s]$$

That is, applying $\lambda x.t$ to s yields the value t where x is interpretated as s. So for example $\lambda x.(x{+}2)$ applied to 3 yields the value 3+2. If the λ operator is *untyped*, then there are no restrictions on what it can take as its values. That is, any function can be applied to any function, including itself.

The full system of rules for untyped λ-calculus can be found in [8], here it is enough to note that these rules validate β-reduction and maintains the admissibility of (λ-Cut) and (λ-Reflexivity). Therefore, on the proof theoretic account of logic, the untyped λ-calculus is a 'logic' of the basic semantic principle (3): i.e. it is a logical syntax of computation.

The system of rules in [8] is 'full' in the sense that it derives all the reductions of what is traditionally studied in books on λ-calculus. But, just as in the case of logic, there is no reason why more rules for new operators cannot be defined, as long as they preserve the structural rules of (4).

λ-calculus is a now well studied formal system developed by Church as a general framework for reasoning on functions (in particular, function application and abstraction). Untyped λ-calculus is a theory of computation in at least two senses: it is the most general and abstract theory of functions available, and it is itself powerful enough to represent all computable functions. In other words, untyped λ-calculus is powerful enough to be able to give an abstract characterisation of any Turing machine. Similarly, untyped λ-calculus can represent any recursive function on the natural numbers, and hence, recursive arithmetic (e.g. see e.g. [13]).[7]

4.2 λ-theories

An additional feature of the system of [8] is that it can reason on λ-theories. A λ-theory is a set of assumptions about what computations are possible. In the case of logic, we can assume the proposition $\neg B$ and then conclude on the basis of this assumption that $A \vee B$ implies A. Similarly with computation, under the assumption that, say, f computes to g, other computations may arise (e.g. that $f(x)$ computes to $g(x)$).

A λ-theory has the computational significance of identifying which computations need to be specially encoded in the software, and which can be assumed in the hardware. For example, a digital computer that does not have an in-built module for arithmetic will need to find some complex representation of numbers and arithmetic functions in order to perform simple arithmetic calculations. But to improve efficiency, modern computers have a special hardware module for doing arithmetic, as far as programming is concerned 3+2 automatically computes to 5 without having to program it in.

[7] So Hacking's comment (page 2 of this paper) was premature, but reasonable as the system of [8] was not known at the time. Although the λ-calculus can represent every recursive function as a λ-term, using proof theoretic means to justify the principle of induction, to obtain full Peano arithmetic, is problematic. For an alternative, formalist, strategy see [7].

For example, a computer can have a representation of the functions $1, 2$ and $\times$ in at least two different different ways. On the one hand it could take them as primitive constants, and have a hardware module that automatically rewrites, say, $1{\times}2$ to 2. Alternatively, the computer could encode $1, 2$ and $\times$ as complex functions so that $2{\times}2$ computes to 4 using the laws computation alone (i.e. in untyped λ-calculus). An example of such an encoding is given by *Church numerals*.[8]

$$\begin{aligned} 1 &\text{ is encoded by } \lambda f.\lambda x.f(x) \\ 2 &\text{ is encoded by } \lambda f.\lambda x.f(f(x)) \\ &\vdots \\ \times &\text{ is encoded by } \lambda x.\lambda y.\lambda f.\big((x(y))(f)\big) \end{aligned}$$

It is then a straightforward matter to show that on this encoding, $\times(2(2))$ computes to 4 (mainly by β-reduction). Notice that on this encoding, $1{\times}2$ computes to 2 purely on the basis of how the relevant functions are represented as λ-terms.

4.3 What is a computer?

We can now give an account of computation that parallels the proof theoretic account of logic. In the case of logic, a 'logical step' could be defined as an inference in accordance with a logical inference rule (for example, inferring A from $A \wedge B$). In the case of computation, a computational step can be defined as a step in accordance with a rule of the λ-calculus, or some extension of it that preserves (4).

This paper proposes that a computer may be characterised as a physical system that implements a λ-theory in (an extension of) untyped λ-calculus. Given the equivalence of the λ-calculus and Turing machines it follows that this definition is compatible with the classical definition of Section 3.2. A program is then a particular λ-term and the computer runs the program by computing on it by means of the rules of untyped λ-calculus and the assumptions of the λ-theory it implements.

But what is it for something to *implement* a λ-theory? An implementation of the λ-calculus is an internal language of functions (both as arguments and values), function application and function abstraction (the λ-operator). Furthermore, these languages must obey the rules of the λ-calculus and the assumptions of the λ-theory. This internal language need not be primitive to the architecture of the computer, the computer's 'word' for, say, the function abstraction λx, may be a complex of states and state changes. All that is required is that if we do view these states as λ-terms then the computer can be seen as following the laws of the λ-calculus.

[8] The Church numerals bear a striking similarity to the Wittgenstein's account of arithmetic in [20, 6.021–6.03]. Wittgenstein suggests that a number, n, is a propositional function that repeats any other propositional function n times. This is similar to the content of a Church numeral within the λ-calculus. Combining this with the background theory of propositional functions Wittgenstein was also trying to develop we get a theory that anticipates the λ-calculus to a substantial degree.

We can put this more formally:

> A physical system implements a λ-theory when its inputs and outputs correspond (one-one) to the language of the λ-calculus such that state y is an output of state x only if the λ-term corresponding to x reduces (or computes) to the λ-term corresponding to y in the λ-theory. A computer is a physical system that implements a λ-theory.[9] (5)

This definition is far from restrictive, for, given a physical system, there is no fact of the matter what its inputs and outputs are. That is, we are free to interpret any states of a system as its inputs and outputs. So the same physical system can be seen as seen as implementing computations in many different λ-theories. Thus a computer may be seen as running many different programs simultaneously. This is not problematic, for we have already discussed in Section 3.2 that computation is relative to an interpretation. Exactly what a computer computes depends not only on it, but on an interpretation of what its states, inputs and outputs are.

So to answer the original question 'what is computation?', we can say that a computation is a sequence of computational steps. A computational step is one in accordance with the rules of a suitable extension of untyped λ-calculus. A computer is then a (physical) implementation of a λ-theory and a program is a λ-term within that theory.[10]

This foundation for computation is entirely hardware independent, λ-theories make no mention of hardware, and so different computational architectures could implement the same λ-theory.

5 Two applications: Turing-Thesis and The Mind

5.1 A new argument for Turing's Thesis

The proof theoretic account of computation provides us with a new basis for 'proving' Turing's Thesis. Actually, we get an argument for what we could call the Church-thesis, meaning the thesis that every computable function is representable in the untyped λ-calculus.

The proposed argument is similar to that of Smith in [18], although it perhaps slightly stronger in that it involves a formal analysis of the concept of computation.

The argument itself is quite simple. It begins with the premise that a fundamental principle to the concept of a computation is the semantic principle (3) given on page 4. That is, the relation of computation that holds between an argument and value adheres, *at least*, to (3). Secondly the argument assumes the proof theoretic method of fixing the meanings of terms (for which some argument has been made in this paper). Given this we can accept the untyped λ-calculus as a governing logic of computable functions. Furthermore, as noted above, all recursive functions can be represented in the untyped λ-calculus.[11] Now, if we can show that no further extension of the λ-calculus is necessary to represent computable functions then we can conclude that the untyped λ-calculus is *the* governing logic of computable functions and so that every computable function can be represented in terms of λ-terms. But, any further proof theoretic extension to the λ-calculus can itself, by a process of Goedel coding, be captured within the λ-calculus (as we can already represent all recursive relations as λ-terms). So we may conclude that computable functions are completely governed, logically, by the laws of the λ-calculus and so are representable as λ-terms.

The above argument makes a crucial assumption, that the intuitive notion of computation is logical in the sense that it is independent of any fact of the matter of how the world actually is. That is why we can assume, on the proof theoretic approach, that that its governing laws must be characterisable as a formal system of inference rules adhering to the structural regularities of (4). This would be disputed by anyone who believes that our notion of computation might be extended by certain empirical findings, or perhaps those who reject the proof-theoretic treatment of what counts as a *logical* inference system. Authors of either type can be found within the literature on *hypercomputation* who would also argue further that they could present (what they claim to be) computable functions, dependent of infinite processes, which are not representable in the λ-calculus.

A detailed discussion of these opposing views is beyond the scope of this paper. However, we can hint here that they need to be seen as being completely contrary to the argument presented here. The argument here depends on a purely finitary proof-theoretic account of a computational law (drawn from the infinitary account of a logical law). If we extend this account to allow for infinitary proof-theoretic justifications of inference rules, then we can generate an argument for a 'hyper'-Turing-thesis. Although its exact status and legitimacy would be a matter of substantial further discussion.

5.2 Computationalism, a new account

Computationalism is the view that the mind is a kind of computer. There are many versions of it, perhaps one of the most modest versions is the idea is that mental processes (usually theorised as physical systems) form a certain referential language on which the mind computes, symbolically. So this version of computationalism combines a simple, perhaps causal, theory of mental reference (or, representation) with an interpretation of mental processes as computations.[12]

If we have the kind of view of computation exemplified by the Fodor quotation of page 3, then the content of mental processes is carried almost entirely by the theory of the reference of a mental statae. This is because the symbolic manipulation envisaged is merely the writing and rewriting of 'symbols' in the language of thought according to some set of instructions. It becomes no surprise then that such a theory struggles to explain mental content. Semantic objections to computationalism, of which Searle's Chinese room is one example, have

[9] 'Only if' and not 'if and only if'. A computer need only be *sound* with respect to the λ-calculus: it need not output everything that could possibly be computed from a given input.

[10] This does not mean that all computer programming should be done using the λ-calculus (which is impractical, on its own, as a programming language). A computer program is merely something that can be interpreted as implementing a λ-theory in the sense of (5).

[11] This is a good 'sanity check' of the formal analysis of computation proposed here, for the recursive functions are indeed intuitively computable.

[12] This is arguably the view expressed by Fodor in his earlier work [3].

intuitive force as the representational theory bears the full load of explaining mental content and, it seems, struggles to do so.

Section 4.3 makes a more subtle account of computation available to computationalists. A computation is the reduction of a λ-term in accordance with a law of computation. Thus a λ-term has *computational content* just virtue of being a λ-term (in the same way that conjunction has logical content virtue of being a logical connective). So if we replace 'term' and 'langauge of thought' with 'λ-term' and 'λ-theory of thought', then we obtain a theory of computationalism where mental processes *can* have content: computational content.

The idea is then as follows. We assume some basic theory of reference, perhaps a simple causal theory as discussed by Fodor in [3], then we postulate that:

> Mental states are referential states (according to some suitable reference theory), furthermore the mind is a computer the states of which are complex λ-expressions. Mental processes are computations on these complex λ-terms in accordance with the laws of computation (which are laws for proof-theoretic reasons). Finally, the content of a mental state is given not merely by its referential content, but also by its computational content. (6)

To exemplify this, return to the arithmetic example of Section 4.2. In that example the functions 2, 4 and $\times$ are encoded as λ-terms such that 2×2 reduces to 4 just in virtue of the structure of the terms. The λ-term 2 therefore has its own independent computational content: part of that content is that when it is applied twice to $\times$ it computes to the term encoding 4 (by the laws of computation). To use a more linguistic example, the proposal is that a concept, say 'cat', is encoded mentally by a λ-term that not only refers cats, but is sufficiently complex that it computes to the λ-term that refers to animals. Thus it can be contained computationally in the content 'cat' that cats are animals.

We can identify the content of a λ-term by the set of all λ-terms to which it computes.[13] Many of the terms that a given λ-term computes to themselves have referential content. So we can gloss (6):

> The mind refers the external world using a language of thought; the terms of this language are complex and are related to each other by the laws of computation; the full content of a term is given by its referential content and the contents of all the terms to which it computes (and many of these terms will also possess referential contents). (7)

Thus we obtain the intuitive picture that the content of a term is grounded in its relations to the contents of other terms. Furthermore we can say in some detail what that relation is, it is the relation of computation (via the computational laws, which can be justified proof theoretically).

Before applying this account to a famous objection to computationalism I wish to hint that the account of mental (computational) content is actually quite close to the empirically motivated 'connectionist' accounts. The reader is referred to [9] where a model-theoretic semantics for λ-terms is developed in terms of a ternary relation on elements of a domain. We can interpret this ternary relation as relating an element, or 'node', of the domain to its inputs and outputs. This type of structure forms the basis of the theory of neural networks, unfortunately the semantic relation between λ-calculus and neural nets is beyond the scope of this paper.

5.3 A return to the Chinese room

We can now return to the Searle's Chinese room example and apply this proof-theoretic theory of computation to it.

Imagine a robot that acts so as to fool everyone into believing it is human and Chinese. There is a strong intuition that that robot must be complex enough to possess mental states, like humans. It is also natural to conclude further that it is the programming that determines the mental content and not the exact physical make-up of the robot. We might then theorise that humans are a particular kind of organic robot that instantiate a very complex programme, and it is virtue of this programming that humans have complex mental content (beyond simple reference). But we have the strong intuition that there is no substantial mental content, or understanding, in Searle's Chinese room, which is merely an implementation of a look-up table. Searle then concludes that since the programming is the same – by assumption, the room also fools everyone into believing it contains a Chinese human – it would follow that humans have no mental content either. Searle then rejects computationalist theories of the mind.

The flaw in this argument is that it assumes that Turing equivalent (indistinguishable by their inputs and outputs) are running the same programs. Although the Chinese room behaves just like a human Chinese speaker, it does not follow that it is running the same type of programming.

But how could computationalists differentiate between a look-up table, Chinese room type of 'speaker' and a real person? The answer, given the theory of computational laws just presented, is now straightforward. A look-up table implements a λ-theory that contains many assumptions (one for each line in the look-up table), and its programs are very simple λ-terms, constants in fact. That is, each input is a constant function, and we have a huge stock of assumptions that specify how to manipulate them. The computational content of these terms is therefore quite light. A more subtle computer, or a person, that really does speak Chinese will implement Chinese as a λ-theory with fewer assumptions, and where computations on Chinese inputs are done on the level of how they are represented, i.e. the λ-terms that are associated (by sameness of reference) with the Chinese words. The terms of such a computer bear significantly greater computational content.

The claim here is that when we have the intuition that a computer does not really 'think' or 'understand' or posses mental content, we are having the intuition about a computer of the look-up table variety. The programs in such a computer are very simple λ-terms that have no computational life independently of a very large set of additional assumptions specifying their behaviour.[14] But a computer need not be programmed that way. According to the proof theoretic

[13] Similarly, we can identify the logical content of a proposition by all the propositions it entails.

[14] This seems implicit in the earlier quotation from Fodor.

account of computation, a computer can implement complex λ-terms and carry out computations purely on the basis of the structure of these terms.

Let us take this response directly to Searle:

> Because programs are defined purely formally or syntactically, and because minds have an intrinsic mental content, it follows immediately that the program itself cannot constitute the mind. The formal syntax of the program does not by itself guarantee the presence of mental contents. I showed this a decade ago in the Chinese room argument. [17]

We can now reply that, on a better model of computation, formal and syntactic systems can have intrinsic content: they can have *intrinsic computational content*. Given this, Searle's argument fails as it presumes there is no suitable notion of content to associate with formal syntax.

Thus we can rescue computationalism from the Chinese room argument. We use our more subtle notion of computation say that mental states are special kinds of computer program, and that although implementing such a program is hardware independent, two computers can be Turing equivalent without implementing the same programs. That is, two computers can yield the same inputs from the same outputs but be running programs with distinct computational contents: one can be an almost computationally contentless look-up table; the other could be a subtle piece of programming, rich in content.

6 Conclusion

The main conclusion of this paper is that there are such things as *laws* of computation and there is a proof-theoretic explanation of what they are: untyped λ-calculus (the general theory of functions). Logic and computation can then be seen as having the same, proof-theoretic, foundations. As this foundation for computable functions is recursive, we can offer a proof-theoretic argument for Turing's thesis, that all computable functions are recursive (or can be emulated by a Turing machine).

Furthermore, given the laws of computation, the notion of computational content then becomes easy to define. The computational content of a term is its computational relations to other terms according to the laws of computation. This in turn gives an extra dimension to computationalism, for mental states can have computational as well as representational content.

This more subtle version of computationalism is better equipped to answer famous objections such as the Chinese room argument. The Chinese room, although indistinguishable from a Chinese speaker, has no mental content as the representations (of Chinese expressions) have no computational content — the computation follows a look-up table, beyond the laws of computation alone. In the case of a typical Chinese speaker, Chinese expressions have computational, as well as representational content, this yields the speaker's mental content. A computationalist theory should seek to identify mental content both with referential and computational content. Whether this more subtle version of computationalism succeeds is for further analysis, however it does not fail on the basis of the Chinese room argument.

REFERENCES

[1] Arnon Avron, 'Multiplicative conjunction and an algebraic meaning of contraction and weakening', *Weakening, forthcoming in the Journal of Symbolic Logic*, (1998).

[2] Robert Brandom, 'Varieties of understanding', in *Reason and Rationality in Natural Science*, ed., Nicholas Rescher, 27, University Press of America, Lanham, (1985).

[3] Jerry Fodor, *The elm and the expert: mentalese and its semantics*, MIT Press, 1994.

[4] Jerry Fodor, *The Mind Doesn't Work That Way*, MIT Press, 2001.

[5] Gottlob Frege, *The Foundations of Arithmetic*, Blackwell, Oxford, 1953. tr. by J. L. Austin.

[6] Gottlob Frege, *The Basic Laws of Arithmetic*, Berkeley: University of California, 1967. M. Furth (trans.).

[7] Michael Gabbay, 'A formalist philosophy of mathematics part I: Arithmetic', *Studia Logica*, **96**(2), 219–238, (2010).

[8] Michael Gabbay, 'A proof-theoretic treatment of λ-reduction with cut-elimination: λ-calculus as a logic programming language', *Journal of Symbolic Logic*, (June 2011).

[9] Michael Gabbay and Murdoch James Gabbay, 'A simple class of kripke-style models in which logic and computation have equal standing', in *LPAR (Dakar)*, eds., Edmund M. Clarke and Andrei Voronkov, volume 6355 of *Lecture Notes in Computer Science*, pp. 231–254. Springer, (2010).

[10] Gerhard Gentzen, 'Untersuchungen über das logische Schliessen', *Mathematische Zeitschrift*, **39**, 176–210 and 405–431, (1934).

[11] Ian Hacking, 'What is logic?', *The Journal of Philosophy*, **76**, 285–319, (1979).

[12] Stevan Harnad, 'What's wrong and right about searle's chinese room argument?', in *Essays on Searle's Chinese Room Argument*, eds., M. Bishop and J. Preston, Oxford University Press, (2001).

[13] J. Roger Hindley and Jonathan P. Seldin, *Lambda-Calculus and Combinators, An Introduction*, Cambridge University Press, 2nd edn., 2008.

[14] S. C. Kleene and J. B. Rosser, 'The inconsistency of certain formal logics', *Annals of Mathematics*, **36**(3), (1935).

[15] Christopher Peacocke, 'Understanding logical constants: a realist's account', *Proceedings of the British Academy*, **73**, (1987).

[16] John Searle, 'Minds, brains, and programs', *Behavioral and Brain Sciences*, **3**, 417–457, (1980).

[17] John Searle, *The rediscovery of mind*, Cambridge, MA: MIT Press, 1992.

[18] Peter Smith, *An Introduction to Godels Theorems*, Cambridge University Press, 2007.

[19] Alan Turing, 'Computing machinery and intelligence', *Mind*, **59**, 433–460, (1950).

[20] L. Wittgenstien, *Tractatus logico-philosophicus, translated by C.K. Ogden*, Routledge & Kegan Paul, 1955.

A mouse in the Chinese room

Etienne B. Roesch[1] and **Slawomir J. Nasuto**[2] and **J. Mark Bishop**[3] and **Matthew Spencer**[4]

Abstract. John Searle's Chinese Room Argument (CRA) pertains to demonstrate that syntax is not sufficient for semantics, and that because computation cannot yield understanding, the computational theory of mind, which equates the mind to an information processing system based on formal computations, fails. These criticisms also contribute to show the inadequacy of the Turing test to demonstrate intelligence. In this paper, we use the CRA, and the debate that emerged from it, to develop a philosophical critique of recent advances in robotics and neuroscience. We describe results from a body of work that contributes to blurring the divide between biological and artificial systems: so-called animats, autonomous robots that are controlled by biological neural tissue, and what may be described as remote-controlled rodents, living animals endowed with augmented abilities provided by external controllers. We argue that, even though at first sight these chimeric systems may seem to escape the CRA, on closer analysis they do not. We conclude by discussing the role of the body-brain dynamics in the processes that give rise to genuine understanding of the world, in line with recent proposals from enactive cognitive science[5].

1 Searle's Chinese Room Argument

1.1 The CRA in a nutshell

Arguably, John Searle's CRA yielded one of the most notorious debates in the history of philosophy of mind [4, 15]. His argument pertains to the strong claim of artificial intelligence, which he coined "Strong AI", of creating a truly intelligent computational device, demonstrating machine understanding [19, p. 417]. This lasting debate has important consequences for cognitive science in general, and the computational theory of mind in particular, which equates the mind to an information processing system based on formal computations [8, 9, 16]. It also shows the inadequacy of any purely behavioural imitation game (e.g., the Turing test) to identify intelligence[6].

In a nutshell, the CRA responds to Schank and Abelson's account of their computer program that was said to simulate the human ability to understand short stories [18]. The program would take as input a formatted version of a story and, using sets of rules, heuristics and scripts, would infer answers to questions posed by an operator. Scripts referred to the detailed description of the stereotypical events unfolding through time in given contexts. Searle gives the example of a story depicting a man entering a restaurant, ordering a hamburger and storming outside the restaurant disappointed. When asked "Did the man eat the hamburger?", the program would unequivocally answer "No, he did not", based on what is expected of these sorts of stories.

Searle aimed to demonstrate that syntax is not sufficient for semantics, and that a rule-based program, such as Schank and Abelson's, will never be able to explain the human ability to genuinely understand a story [19]. To this end, he described a thought-experiment in which he was locked in a room, and provided with notes written in Chinese. Searle does not speak a word of Chinese, and to him "Chinese writing is just so many meaningless squiggles" (p. 418). With him in the room can be found a rulebook written in English, and a second batch of Chinese squoggles. The rulebook describes the squoggles that should be produced in answer to the squiggles that are passed to him through a crack in the door. After a little while, he argues, a naive, Chinese speaking observer witnessing the scene would wrongfully assume that whoever is locked in that room genuinely understands Chinese, for they would seem to provide adequate textual responses and exhibit real Chinese mannerisms in their linguistic style. Searle goes on to demonstrate that, even though he could master the rulebook perfectly, he would still not understand a word of Chinese.

According to Bishop [2],"the central claim of the CRA is that computations alone cannot in principle give rise to understanding, and that therefore computational theories of mind cannot fully explain human cognition. [...] And yet it is clear that Searle believes that there is no barrier in principle to the notion that a machine can think and understand; [...] Searle explicitly states, in answer to the question 'Can a machine think?', that 'the answer is, obviously, yes. We are precisely such machines'" (p. 47).

Searle's "intuition pump", a term coined by Dennett [6], provoked an intense reaction in the AI community who attempted, but arguably failed, to demonstrate that the CRA was wrong. Amongst these criticisms, which Searle anticipated in the original exposition of the CRA, are what he identified as the "Systems reply", the "Robot reply" and "the Brain Simulator reply". Searle takes these replies to the CRA very seriously, presciently anticipating several key turns in recent cognitive robotics, AI and cognitive science.

In what follows, we will briefly review the above replies, before introducing a number of successes in a new branch of robotics that contributes to blurring the divide between biological and artificial

[1] Goldsmiths, Univ. of London, UK / Univ. of Reading, Reading, UK, contact@etienneroes.ch
[2] Univ. Reading, of Reading, UK, s.j.nasuto@reading.ac.uk
[3] Goldsmiths, Univ. of London, UK, m.bishop@gold.ac.uk
[4] Univ. Reading, of Reading, UK, matthew.spencer@pgr.reading.ac.uk

[5] In this work the term enactive cognitive science will be used to delineate theoretical approaches to cognition that emphasise perception as action encompassing, for example, Gibson's "ecological approach"; Varela, Thompson and Rosch's "embodied mind"; Nöe's "action as perception" and O'Regan and Noë's "sensorimotor account of vision".

[6] In what has become known as the standard interpretation of the Turing test a human interrogator, interacting with two respondents via text alone, has to determine which of the responses is being generated by a suitably programmed computer and which is being generated by a human; if the interrogator cannot reliably do this then the computer is deemed to have passed the Turing test.

systems. We aim to use these examples to articulate a response to current trends in cognitive robotics in line with Searle's original line of thought as he espoused in the CRA.

1.2 The Systems reply

The Systems reply originated from researchers who took a bird's-eye view of Searle's thought-experiment. To them, understanding, if any, does not lie within Searle but within the system as a whole; it is the room plus Searle, plus the rulebook, plus the squiggles and squoggles, which exhibits responses that are perceived as genuinely Chinese. Searle responds to this line of thoughts by pointing out that, even if he could memorise the rulebook and all the paper squiggles and squoggles, and hence iconographically interact with Chinese people directly, laboriously following the instructions of the rulebook as memorised, he would still not understand a word of Chinese. The Systems reply has deep ramifications, because it links the claim of AI to Second Order Cybernetics whereby the Observer of the system plays an active role in the engagement of the system with the world. Searle, however, argues that even this stance only hides the problem at a different level; understanding, he adds, does not lie in the mere syntactic exchange of input and output symbols–doing so, however coherently, can never warrant instantiation of genuine understanding of intentional content.

1.3 The Robot reply

The Robot reply acknowledges that understanding requires some degree of interaction with the world. Proponents of this position extend the CRA to a robot that interacts with the world through actuators, and perceives it through sensors. Using scripts similar to Searle's rulebook, a computer decides the appropriate symbolic response (squoggles) to the symbolic descriptions of the world (squiggles) being presented. Surely in this case the robot could be said to understand Chinese, they infer. However, making what is by now a well rehearsed move, Searle once again argues that as both the squiggles and squoggles, are merely uninterpreted symbols, a series of binary digits, they would still remain meaningless to him, even if he sat in place of the central computer. In other words the situation Searle describes highlights the possibility of a zombie robot; one that would always fail to obtain the intimate connection with the world required to give rise to a sense of intentionality, understanding and meaning.

1.4 The Brain Simulator reply

The third reply assumes a computer model of the neural mechanisms at play in the brain of a native Chinese speaker, when they understand stories in Chinese and provide answers about these stories. Advocates of this side of the debate asserts that denying understanding to this computer model equates to denying understanding to native Chinese speakers. Searle responds to this proposal by suggesting a replacement of the neurons and synapses with a complex functional analogue made from an interconnection of water-pipes and valves, each of which he would activate according to a new rulebook upon receiving a specific series of squiggles as input. Perhaps not surprisingly, Searle once again concludes that, to him, the Chinese squiggles would remain meaningless.

2 Hybrid systems and levels of embodiments

The replies to Searle's thought-experiment describe situations that are both relevant and conceivable: each situation emphasises particular perspectives of the CRA, and could give rise to further investigation in the form of actual physical/biological experiments with tangible implementations. Today's proponents of so-called embodied AI, a field now known as cognitive robotics, in fact take at least some of Searle's comments very seriously, and argue that endowing robots with (cognitive) abilities to reason about the world will demonstrate intelligent behaviour and, consequently, exhibit a genuine understanding of the world. This work has employed extremely varied strategies, the success of which, however, remains debatable [17].

Interestingly, a fourth line of reply to the CRA involves the combination of the previous three replies. In this particular situation, one assumes the examination of a robot in the world, operated by a synthetic brain modelled after a native Chinese speaker. Searle agrees with the contenders of this line of thought: "I entirely agree that in such a case we would find it rational and indeed irresistible to accept the hypothesis that the robot had intentionality, *as long as we knew nothing more about it.*." [19, p. 421; our emphasis].

In the present paper, we posit that recent technological advances, which contribute to blurring the divide between biological and artificial systems, may serve as a vehicle to push this examination further. In particular, we focus on so-called animats [23, 11], autonomous robots that are controlled by biological neural tissue, and what may be described as remote-controlled rodents, living animals endowed with augmented abilities provided by artificial controllers. These two chimera can be seen as the two sides of the same coin and, we argue come a step closer to the physical realisation of the well known "brain-in-a-vat" thought-experiment, cousin to the CRA.

2.1 Animats

Recently, one of the co-authors led a team at the University of Reading that successfully developed an autonomous robot controlled by cultured living neural cells [22, 21]. The "brain" of the system consisted of a cultured network of thousands of neurons, sliced from the cortical tissue of foetal rats, and grown on an array of electrodes that permits both recording and electrical stimulation. As a result of the procedure, the connections between the neurons are lost, but within a short period of time, new connections spontaneously form, and neurons start engaging in communication. The activity grows over the subsequent weeks into bursts of activity that spread over the entire culture until maturation (about 1 month after seeding). The resulting activity was then used to control the actuators of a small wheeled robot and, closing the loop of the system, the signal registered by the robot's sensors was being fed back to the cultured neurons in the form of brief electrical impulses. This platform demonstrated simple obstacle avoidance behaviours[7], analogous to a simple Braitenberg vehicle, and the *a posteriori* analysis of the cultures showed functional connectivity, as well computational and biophysical properties similar to that of intact brains.

2.2 Programming rodents

In recent years, successes in implant technology gave rise to functional hybrid systems integrating artifacts with the nervous systems of living organisms. Efforts in this direction are motivated by the

[7] See `www.youtube.com/watch?v=1-0eZytv6Qk`.

creation of prostheses, e.g. cochlear [3] or retinal [24, 10] implants, and are now moving beyond augmenting sensory modalities towards interfacing directly with the brain through deep brain stimulation. This technique involves implanting tiny electrodes in nuclei of the brain, permitting the recording and stimulation of local neurons. It is an approved clinical technique for the treatment of many neurological disorders in humans [7, 14]. Before reaching this stage, however, extensive testing has to be performed on seemingly simpler brains, like that of rodents. Berger et al., for instance, successfully demonstrated implants that replaced a rat's hippocampus during a spatial memory task: when the device was inactivated, the animal failed the behavioural task; its performance was restored when the device was switched back on [1]. Another example comes from John Chapin's group, whose implant coupled reward and sensory processing areas in an operant conditioning procedure to train the rat to respond behaviourally to particular tactile stimulations [20]. Upon several days of training, the 'programmed rodent' was able to follow commands, henceforth behaving in ways much similar to a remote-controlled animal[8]. See also Gradinaru et al [13], who used optogenetic techniques to stimulate neurons selectively, inducing motor behaviour without requiring conditioning.

2.3 From an "intuition pump" to the physical realisation of thought-experiments

Does our animat, which successfully avoids obstacle, genuinely understand it is facing a wall? Can the remote-controlled mouse, which turns left in infinite loops as long as the device is switched on, be said to genuinely understand what it is doing, and why? How about the remote-controlled rat that blindly follows motor commands, does it understand the input it receives?

The situations we described, we posit, push Searle's CRA a little bit further, permitting the philosophical exploration of the fourth line of reply to the CRA, that of the robot endowed with a brain much alike to a biological brain. In these situations, both the animats and what we called remote-controlled rodents experiments assume some degrees of embodiment and relatedness to the workings of biological brains. A systems view would thus legitimately raise the question of the intrinsic understanding that these chimera might exhibit of their "personal predicament".

Cosmelli and Thompson already paved the way for this line of thought in an attempt to formulate a response to the "brain-in-a-vat" thought-experiment, a cousin to the CRA, about consciousness [5]. In this experiment, the reader is invited to imagine a brain floating "in a life-sustaining vat of liquid nutrients" and connected to "a supercomputer that would stimulate it with electrical impulses exactly like those it normally receives when embodied" (p. 361). Cosmelly and Thompson use this thought-experiment to explore the role of the body in the definition of consciousness. Notably, they pose that a functional body is required to support consciousness, and that such body needs to be "a self-regulating system comprising its own internal, homeodynamic processes and capable of of sensorimotor coupling with the outside world" (p. 363), a conception at the heart of the enactive approach to cognitive science, and conclude that "consciousness is a function of life-regulation processes involving dense couplings between neuronal and extraneuronal systems, rather than a function of neural systems alone" (p. 379).

We argue that, even though their interest in this thought-experiment lies in the defining features of consciousness, their argument might as well apply to intentionality. In fact, as an analogue of Cosmelli and Thompson's "brain-in-a-vat", we suggest that the impoverished notion of a "body" that serves the animat equally offers no hope for anything more than mere sensory-motor coupling to arise.

The lack of proper embodiment is, however, only part of the problem, as demonstrated by the remote-controlled rodent experiments. In these cases, the chimera is constituted by a fully functional, living body that is endowed with an artificial device that transmits electrical signals to the brain, to which it responds. Behaviorally, the remote-controlled animals seem to lose something more than "just" free-will and volition. We suggest that, even though the animals still possess a fully functional body and, arguably, a functioning brain, the fact that it receives alien commands does not warrant a genuine understanding of what is going on. In other words, the animal's brain receives foreign input that, at best, may resemble drug-induced decontextualised hallucinations–mere uninterpreted symbols/squiggles and squoggles–,which, we argue, would remain meaningless despite how accustomed the animal may become to this new mode of ownerless functioning. The situation is analogous to the alien hand syndrome - in such patients, for example, their arm seemingly performs actions not their volition or under their control (in fact, often against their will). Such patients do not accrue any meaning of why their arm acted in this way, albeit they can see (and hence comprehend) the actions in the same way as any other observer; in this sense they are 'external observers' of their own limb(s).

It is thus clear that genuinely experiencing intentional states, in the process of understanding the world, requires both a fully functional brain and a fully functional body; deporting the question of the requirements for genuine understanding to the defining features of the process whereby the brain and body interact with the world. Without this, neither our animats, nor the remote-controlled rodents experiments can escape Searle's CRA. This argument lends support to discussions of the properties grounding the agent-environment system. Fröse and Ziemke, for instance, discuss the foundational role of constitutive autonomy and adaptivity [12] for agency and sense making, and their consequences for the design of embodied AI.

3 Summary and conclusion

In this paper, we based a philosophical examination of the requirements for genuine understanding and intentionality on extensions to John Searle's Chinese Room Argument against strong AI. Our deployment of Searle's "intuition pump" to recent advances in robotics and neuroscience shows it continues to have force against the most recent developments in robotics and bio-machine hybrids. Specifically, we examined how two new scenarios fare in light of the CRA's typical replies from the AI community. We focused on so-called "animats", autonomous robots that are controlled by biological neural tissue, and what may be described as "remote-controlled rodents", living animals endowed with augmented abilities provided by artificial controllers. These two chimera can be seen as the two sides of the same coin, and herein we have demonstrated that neither of these systems can be said to exhibit any genuine understanding of the world.

In addition, we argue that current efforts in cognitive robotics, to endow robots with abilities to represent the world and reason about it, are limited. In line with the rise of enactive cognitive science - which proposes an enlarged perspective that includes the closed-loop interactions of a life-regulated body-brain dynamical system with an evolving world - we deem inappropriate cognitivism and its concomitant computational theory of mind, and instead emphasise the

[8] See `www.youtube.com/watch?v=D5u2IWFNFDE`.

role of foundational processes such as autonomy, exploration, autopoiesis and social embedded-ness, in giving rise to a genuine understanding of our lived world.

ACKNOWLEDGEMENTS

We would like to thank Dr. Tom Fröse and an anonymous reviewer for their comments, which helped improve this paper. Please note that, in the context of the 'Computing, Philosophy and the Question of Bio-Machine Hybrids: 5th AISB Symposium on Computing and Philosophy', part of the 2012 Turing centenary AISB/IACAP World Congress, this paper specifically re-examines arguments first discussed by Nasuto and Bishop in '*Of (zombie) mice and animats*', and presented at the 2011 PT-AI conference, Thessaloniki. C.f. Nasuto, Slawomir J. and Bishop, J. Mark (2012), '*Of (zombie) mice and animats*', in Vincent C. Müller (ed.), *Theory and Philosophy of Artificial Intelligence*, (forthcoming, SAPERE; Berlin: Springer).

REFERENCES

[1] T W Berger, R E Hampson, D Song, A Goonawardena, V Z Marmarelis, and S A Deadwyler, 'A cortical neural prosthesis for restoring and enhancing memory', *Journal of Neural Engineering*, **8**(4), 046017, (2011).

[2] J Mark Bishop, 'A view inside the chinese room', *Philosopher*, **28**(4), 47–51, (2004).

[3] W Blake, *Cochlear Implants: Principles and Practices*, Lippincott Williams & Wilkins, Philadelphia, 2000.

[4] David Cole, 'The chinese room argument', in *The Stanford Encyclopedia of Philosophy*, ed., Edward N Zalta, Stanford University, winter 2009 edn., (2009).

[5] D Cosmelli and E Thompson, 'Embodiment or envatment? reflections on the bodily basis of consciousness', in *Enaction: Towards a New Paradigm for Cognitive Science*, eds., John Stewart, Olivier Gapenne, , and Ezequiel di Paolo, MIT Press, (2011).

[6] Daniel Dennett, *Consciousness Explained*, The Penguin Press, Allen Lane, 1991.

[7] J J Fins, 'Deep brain stimulation', in *Encyclopedia of Bioethics*, ed., S G Post, volume 2, 629–634, MacMillan Reference, New York, 3rd edn., (2004).

[8] Jerry A Fodor, *The Language of Thought*, Harvard University Press, 1975.

[9] Jerry A Fodor, *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, MIT Press, 1987.

[10] A Fornos, J Sommerhalder, and M Pelizzone, 'Reading with a simulated 60-channel implant', *Frontiers in Neuroscience*, **5**, 57, (2011).

[11] Stan Franklin, *Artificial Minds*, MIT Press, March 1997.

[12] Tom Froese and Tom Ziemke, 'Enactive artificial intelligence: Investigating the systemic organization of life and mind', *Artificial Intelligence*, **173**(3–4), 466–500, (March 2009).

[13] V Gradinaru, K R Thompson, F Zhang, M Mogri, K Kay, M B Schneider, and K Deisseroth, 'Targeting and readout strategies for fast optical neural control in vitro and in vivo.', *J Neurosci*, **26:27**(52), 14231–14238, (2007).

[14] M L Kringelbach, Jenkinson N, S L F Owen, and T Z Aziz, 'Translational principles of deep brain stimulation', *Nature Reviews Neuroscience*, **8**, 623–635, (2007).

[15] *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, eds., John Preston and J Mark Bishop, Oxford University Press, New York, 2002.

[16] Hillary Putnam, *Representation and Reality*, MIT Press, 1988.

[17] Etienne B Roesch, 'A critical review of classical computational approaches to cognitive robotics: Case study for theories of cognition.', in *The Hand: An Organ Of The Mind*, ed., Z Radman, MIT Press, Cambridge, USA, (in press).

[18] R C Schank and R P Abelson, *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*, Erlbaum, Hilsdale NJ, 1977.

[19] John Searle, 'Minds, brains, and programs', *Behavioral and Brain Sciences*, **3**, 417–457, (1980).

[20] S K Talwar, S Xu, E Hawley, S Weiss, K Moxon, and Chapin J, 'Rat navigation guided by remote control', *Nature*, **417**, 37–38, (2002).

[21] K Warwick, S J Nasuto, V M Becerra, and B J Whalley, 'Experiments with an in-vitro robot brain', in *Instinctive Computing*, ed., Yang Cai, volume 5987 of *Lecture Notes in Artificial Intelligence*, Springer, (2010).

[22] K Warwick, D Xydas, S J Nasuto, V M Becerra, M W Hammond, J H Downes, S Marshall, and B J Whalley, 'Controlling a mobile robot with a biological brain', *Defence Science Journal*, **60**(1), 5–14, (2010).

[23] S W Wilson, 'Knowledge growth in an artificial animal', in *First International Conference on Genetic Algorithms and Their Applications*, ed., J J Grefenstette, 16–23, Lawrence Erlbaum Associates, Hillsdale, NJ, (1985).

[24] E Zrenner, 'Will retinal implants restore vision?', *Science*, **295**, 1022–1025, (2002).

Implementing Turing Machines in Dynamic Field Architectures

Peter beim Graben[1] and **Roland Potthast**[2]

Abstract. Cognitive computation, such as e.g. language processing, is conventionally regarded as Turing computation, and Turing machines can be uniquely implemented as nonlinear dynamical systems using generalized shifts and subsequent Gödel encoding of the symbolic repertoire. The resulting nonlinear dynamical automata (NDA) are piecewise affine-linear maps acting on the unit square that is partitioned into rectangular domains. Iterating a single point, i.e. a microstate, by the dynamics yields a trajectory of, in principle, infinitely many points scattered through phase space. Therefore, the NDAs microstate dynamics does not necessarily terminate in contrast to its counterpart, the symbolic dynamics obtained from the rectangular partition. In order to regain the proper symbolic interpretation, one has to prepare ensembles of randomly distributed microstates with rectangular supports. Only the resulting macrostate evolution corresponds then to the original Turing machine computation. However, the introduction of random initial conditions into a deterministic dynamics is not really satisfactory. As a possible solution for this problem we suggest a change of perspective. Instead of looking at point dynamics in phase space, we consider functional dynamics of probability distributions functions (p.d.f.s) over phase space. This is generally described by a Frobenius-Perron integral transformation that can be regarded as a neural field equation over the unit square as feature space of a dynamic field theory (DFT). Solving the Frobenius-Perron equation, yields that uniform p.d.f.s with rectangular support are mapped onto uniform p.d.f.s with rectangular support, again. Thus, the symbolically meaningful NDA macrostate dynamics becomes represented by iterated function dynamics in DFT; hence we call the resulting representation *dynamic field automata*.

1 INTRODUCTION

According to the central paradigm of classical cognitive science and to the Church-Turing thesis of computation theory (cf., e.g., [2, 13, 27, 33]), cognitive processes are essentially rule-based manipulations of discrete symbols in discrete time that can be carried out by Turing machines. On the other hand, cognitive and computational neuroscience increasingly provide experimental and theoretical evidence, how cognitive processes might be implemented by neural networks in the brain.

The crucial question, how to bridge the gap, how to realize a Turing machine [33] by state and time continuous dynamical systems has been hotly debated by "computationalists" (such as Fodor and Pylyshyn [8]) and "dynamicists" (such as Smolensky [30]) over the last decades. While computationalists argued that dynamical systems, such as neural networks, and symbolic architectures were either incompatible to each other, or the former were mere implementations of the latter, dynamicists have retorted that neural networks could be incompatible with symbolic architectures because the latter cannot be implementations of the former; see [9, 32] for discussion.

Moore [19, 20] has proven that a Turing machine can be mapped onto a generalized shift as a generalization of symbolic dynamics [17], which in turn becomes represented by a piecewise affine-linear map at the unit square using Gödel encoding and symbologram reconstruction [6, 14]. These *nonlinear dynamical automata* have been studied and further developed by [10, 11]. Using a similar representation of the machine tape but a localist one of the machine's control states, Siegelmann and Sontag have proven that a Turing machine can be realized as a recurrent neural network with rational synaptic weights [29]. Along a different vain, deploying sequential cascaded networks, Pollack [23] and later Moore [21] and Tabor [31, 32] introduced and further generalized *dynamical automata* as nonautonomous dynamical systems (see [12] for a unified treatment of these different approaches).

Inspired by population codes studied in neuroscience, Schöner and co-workers devised *dynamic field theory* as a framework for cognitive architectures and embodied cognition where symbolic representations correspond to regions in abstract feature spaces (e.g. the visual field, color space, limb angle spaces) [7, 26]. Because dynamic field theory relies upon the same dynamical equations as *neural field theory* investigated in theoretical neuroscience [1, 34], one often speaks also about *dynamic neural fields* in this context.

In this communication we unify the abovementioned approaches. Starting from a nonlinear dynamical automaton as point dynamics in phase space in Sec. 2, which has bears interpretational peculiarities, we consider uniform probability distributions evolving in function space in Sec. 3. There we prove the central theorem of our proposal, that uniform distributions with rectangular support are mapped onto uniform distributions with rectangular support by the underlying NDA dynamics. Therefore, the corresponding dynamic field, implementing a Turing machine, shall be referred to as *dynamic field automaton*. In the concluding Sec. 4 we discuss possible generalizations and advances of our approach. Additionally, we point out that symbolic computation in a dynamic field automaton can be interpreted in terms of contextual emergence [3–5].

2 NONLINEAR DYNAMICAL AUTOMATA

A nonlinear dynamical automaton (NDA: [10–12]) is a triple $M_{NDA} = (X, \mathcal{P}, \Phi)$ where (X, Φ) is a time-discrete dynamical system with phase space $X = [0, 1]^2 \subset \mathbb{R}^2$, the unit square, and

[1] Institut für Deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin, email: peter.beim.graben@hu-berlin.de.
[2] Dept. of Mathematics and Statistics, University of Reading, email: r.w.e.potthast@reading.ac.uk.

flow $\Phi : X \to X$. $\mathcal{P} = \{D_\nu | \nu = (i,j), 1 \leq i \leq m, 1 \leq j \leq n, m, n \in \mathbb{N}\}$ is a rectangular partition of X into pairwise disjoint sets, $D_\nu \cap D_\mu = \emptyset$ for $\nu \neq \mu$, covering the whole phase space $X = \bigcup_\nu D_\nu$, such that $D_\nu = I_i \times J_j$ with real intervals $I_i, J_j \subset [0,1]$ for each bi-index $\nu = (i,j)$. Moreover, the cells D_ν are the domains of the branches of Φ which is a piecewise affine-linear map

$$\Phi(\boldsymbol{x}) = \begin{pmatrix} a_x^\nu \\ a_y^\nu \end{pmatrix} + \begin{pmatrix} \lambda_x^\nu & 0 \\ 0 & \lambda_y^\nu \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} , \tag{1}$$

when $\boldsymbol{x} = (x,y)^T \in D_\nu$. The vectors $(a_x^\nu, a_y^\nu)^T \in \mathbb{R}^2$ characterize parallel translations, while the matrix coefficients $\lambda_x^\nu, \lambda_y^\nu \in \mathbb{R}_0^+$ mediate either stretchings ($\lambda > 1$), squeezings ($\lambda < 1$), or identities ($\lambda = 1$) along the x- and y-axes, respectively.

The NDA's dynamics, obtained by iterating an orbit $\{\boldsymbol{x}_t \in X | t \in \mathbb{N}_0\}$ from initial condition $\boldsymbol{x}_0$ through

$$\boldsymbol{x}_{t+1} = \Phi\left(\boldsymbol{x}_t\right) \tag{2}$$

describes a symbolic computation by means of a generalized shift [19, 20] when subjected to the coarse-graining $\mathcal{P}$. To this end, one considers the set of bi-infinite, "dotted" symbolic sequences

$$s = \ldots a_{i_{-3}} a_{i_{-2}} a_{i_{-1}} . a_{i_0} a_{i_1} a_{i_2} \ldots \tag{3}$$

with symbols $a_{i_k} \in \mathbf{A}$ taken from a finite set, an alphabet $\mathbf{A}$. In Eq. (3) the dot denotes the observation time $t = 0$ such that the symbol right to the dot, a_{i_0}, displays the current state, dissecting the string s into two one-sided infinite strings $s = (s'_L, s_R)$ with $s'_L = a_{i_{-1}} a_{i_{-2}} a_{i_{-3}} \ldots$ as the left-hand part in reversed order and $s_R = a_{i_0} a_{i_1} a_{i_2} \ldots$ as the right-hand part. Applying a Gödel encoding

$$\begin{aligned} x &= \psi(s'_L) = \sum_{k=1}^{\infty} \psi(a_{i_{-k}}) b_L^{-k} \\ y &= \psi(s_R) = \sum_{k=0}^{\infty} \psi(a_{i_k}) b_R^{-k-1} \end{aligned} \tag{4}$$

to the pair $s = (s'_L, s_R)$, where $\psi(a_j) \in \mathbb{N}_0$ is an integer Gödel number for symbol $a_j \in \mathbf{A}$ and $b_L, b_R \in \mathbb{N}$ are the numbers of symbols that could appear either in s_L or in s_R, respectively, yields the so-called symbol plane or symbologram representation $(x,y)^T$ of s in the unit square X [6, 14].

A generalized shift emulating a Turing machine[3] is a pair $M_{GS} = (\mathbf{A}^{\mathbb{Z}}, \Psi)$ where $\mathbf{A}^{\mathbb{Z}}$ is the space of bi-infinite, dotted sequences with $s \in \mathbf{A}^{\mathbb{Z}}$ and $\Psi : \mathbf{A}^{\mathbb{Z}} \to \mathbf{A}^{\mathbb{Z}}$ is given as

$$\Psi(s) = \sigma^{F(s)}(s \oplus G(s)) \tag{5}$$

with

$$F : \mathbf{A}^{\mathbb{Z}} \to \mathbb{Z} \tag{6}$$

$$G : \mathbf{A}^{\mathbb{Z}} \to \mathbf{A}^e , \tag{7}$$

where $\sigma : \mathbf{A}^{\mathbb{Z}} \to \mathbf{A}^{\mathbb{Z}}$ is the usual left-shift from symbolic dynamics [17], $F(s) = l$ dictates a number of shifts to the right ($l < 0$), to the left ($l > 0$) or no shift at all ($l = 0$), $G(s)$ is a word w' of length $e \in \mathbb{N}$ in the domain of effect (DoE) replacing the content $w \in \mathbf{A}^d$, which is a word of length $d \in \mathbb{N}$, in the domain of dependence (DoD) of s, and $s \oplus G(s)$ denotes this replacement function.

From a generalized shift M_{GS} with DoD of length d an NDA M_{NDA} can be constructed as follows: In the Gödel encoding (4) the word contained in the DoD at the left-hand-side of the dot, partitions the x-axis of the symbologram into intervals I_i, while the word contained in the DoD at the right-hand-side of the dot partitions its y-axis into intervals J_j, such that the rectangle $D_\nu = I_i \times J_j$ ($\nu = (i,j)$) becomes the image of the DoD. Moore [19, 20] has proven that the map Ψ is then represented by a piecewise affine-linear (yet, globally nonlinear) map Φ with branches at D_ν.

In general, a Turing machine has a distinguished blank symbol, $\sqcup$ delimiting the machine tape and also some distinguished final states indicating termination of a computation [13]. If there are no final states, the automaton is said to terminate with empty tape $s = \sqcup^\infty . \sqcup^\infty$. By mapping $\psi(\sqcup) = 0$ through the Gödel encoding, the terminating state becomes a fixed point attractor $(0,0)^T \in X$ in the symbologram representation. Moreover, sequences of finite length are then described by pairs of rational numbers by virtue of Eq. (4). Therefore, NDA Turing machine computation becomes essentially rational dynamics.

In the framework of generalized shifts and nonlinear dynamical automata, however, another solution appears to be more appropriate for at least three important reasons: Firstly, Siegelmann [28] further generalized generalized shifts to so-called analog shifts, where the DoE e in Eq. (7) could be infinity (e.g. by replacing the finite word w in the DoD by the infinite binary representation of π). Secondly, the NDA representation of a generalized shift should preserve structural relationships of the symbolic description, such as the word semigroup property of strings. Beim Graben et al. [11] have shown that a representation of finite strings by means of equivalence classes of infinite strings, the so-called cylinder sets in symbolic dynamics [18] lead to monoid homomorphisms from symbolic sequences to the symbologram representation. Then, the empty word ε, the neutral element of the word semigroup, is represented by the unit interval $[0,1]$ of real numbers. And thirdly, beim Graben et al. [10] combined NDAs with dynamical recognizers [21, 23, 31] to describe interactive computing where symbols from an information stream were represented as operators on the symbologram phase space of an NDA. There, a similar semigroup representation theorem holds.

For these reasons, we briefly recapitulate the cylinder set approach here. In symbolic dynamics, a cylinder set is a subset of the space $\mathbf{A}^{\mathbb{Z}}$ of bi-infinite sequences from an alphabet $\mathbf{A}$ that agree in a particular building block of length $n \in \mathbb{N}$ from a particular instance of time $t \in \mathbb{Z}$, i.e.

$$\begin{aligned} C(n,t) &= [a_{i_1}, \ldots, a_{i_n}]_t \\ &= \{s \in \mathbf{A}^{\mathbb{Z}} \,|\, s_{t+k-1} = a_{i_k}, \quad k = 1, \ldots, n\} \end{aligned} \tag{8}$$

is called n-cylinder at time t. When now $t < 0, n > |t| + 1$ the cylinder contains the dotted word $w = s_{-1} . s_0$ and can therefore be decomposed into a pair of cylinders $(C'(|t|,t), C(|t| + n - 1, 0))$ where C' denotes reversed order of the defining strings again. In the Gödel encoding (4) each cylinder has a lower and an upper bound, given by the Gödel numbers 0 and $b_L - 1$, $b_R - 1$, respectively. Then

$$\begin{aligned} \inf(\psi(C'(|t|,t))) &= \psi(a_{i_{|t|}}, \ldots, a_{i_1}) \\ \sup(\psi(C'(|t|,t))) &= \psi(a_{i_{|t|}}, \ldots, a_{i_1}) + b_L^{-|t|} \\ \inf(\psi(C(|t| + n - 1, 0))) &= \psi(a_{i_{|t|+1}}, \ldots, a_{i_n}) \\ \sup(\psi(C(|t| + n - 1, 0))) &= \psi(a_{i_{|t|+1}}, \ldots, a_{i_n}) + b_R^{-|t|-n+1} , \end{aligned}$$

[3] A generalized shift becomes a Turing machine by interpreting $a_{i_{-1}}$ as the current tape symbol underneath the head and a_{i_0} as the current control state q. Then the remainder of s_L is the tape left to the head and the remainder of s_R is the tape right to the head. The DoD is the word $w = a_{i_{-1}} . a_{i_0}$ of length $d = 2$.

where the suprema have been evaluated by means of geometric series. Thereby, each part cylinder C is mapped onto a real interval $[\inf(C), \sup(C)] \subset [0,1]$ and the complete cylinder $C(n,t)$ onto the Cartesian product of intervals $R = I \times J \subset [0,1]^2$, i.e. onto a rectangle in unit square. In particular, the empty cylinder, corresponding to the empty tape $\varepsilon.\varepsilon$ is represented by the complete phase space $X = [0,1]^2$.

Fixing the prefixes of both part cylinders and allowing for random symbolic continuation beyond the defining building blocks, results in a cloud of randomly scattered points across a rectangle R in the symbologram. These rectangles are consistent with the symbol processing dynamics of the NDA, while individual points $\boldsymbol{x} \in [0,1]^2$ no longer have an immediate symbolic interpretation. Therefore, we refer to arbitrary rectangles $R \in [0,1]^2$ as to NDA macrostates, distinguishing them from NDA microstates $\boldsymbol{x}$ of the underlying dynamical system. In other words, the symbolically meaningful macrostates are emergent on the microscopic NDA dynamics. We discuss in Sec. 4 how a particular concept, called contextual emergence, could describe this phenomenon [3–5].

3 DYNAMIC FIELD AUTOMATA

From a conceptional point of view it does not seem very satisfactory to include such a kind of stochasticity into a deterministic dynamical system. However, as we shall demonstrate in this section, this apparent defect could be easily remedied by a change of perspective. Instead of iterating clouds of randomly prepared initial conditions according to a deterministic dynamics, one could also study the deterministic dynamics of probability measures over phase space. At this higher level of description, introduced by Koopman et al. [15,16] into theoretical physics, the point dynamics in phase space is replaced by functional dynamics in Banach or Hilbert spaces. This approach has its counterpart in neural [1, 34] and dynamic field theory [7, 26] in theoretical neuroscience.

In dynamical system theory the abovementioned approach is derived from the conservation of probability as expressed by a Frobenius-Perron equation [22]

$$\rho(\boldsymbol{x},t) = \int_X \delta(\boldsymbol{x} - \Phi^{t-t'}(\boldsymbol{x}'))\rho(\boldsymbol{x}',t')\,\mathrm{d}\boldsymbol{x}' \,, \tag{9}$$

where $\rho(\boldsymbol{x},t)$ denotes a probability density function over the phase space X at time t of a dynamical system, $\Phi^t : X \to X$ refers to either a continuous-time ($t \in \mathbb{R}_0^+$) or discrete-time ($t \in \mathbb{N}_0$) flow and the integral over the delta function expresses the probability summation of alternative trajectories all leading into the same state $\boldsymbol{x}$ at time t.

3.1 Temporal Evolution

In the case of an NDA, the flow is discrete and piecewise affine-linear on the domains D_ν as given by Eq. (1). As initial probability distribution densities $\rho(\boldsymbol{x},0)$ we consider uniform distributions with rectangular support $R_0 \subset X$, corresponding to an initial NDA macrostate,

$$u(\boldsymbol{x},0) = \frac{1}{|R_0|}\chi_{R_0}(\boldsymbol{x}) \,, \tag{10}$$

where $|R_0| = \mathrm{vol}(R_0)$ is the "volume" (actually the area) of R_0 and

$$\chi_A(\boldsymbol{x}) = \begin{cases} 0 & : \quad \boldsymbol{x} \notin A \\ 1 & : \quad \boldsymbol{x} \in A \end{cases} \tag{11}$$

is the characteristic function for a set $A \subset X$. A crucial requirement for these distributions is that they must be consistent with the partition $\mathcal{P}$ of the NDA, i.e. there must be a bi-index $\nu = (i,j)$ such that the support $R_0 \subset D_\nu$.

Inserting (10) into the Frobenius-Perron equation (9) yields for one iteration

$$u(\boldsymbol{x},t+1) = \int_X \delta(\boldsymbol{x} - \Phi(\boldsymbol{x}'))u(\boldsymbol{x}',t)\,\mathrm{d}\boldsymbol{x}' \,. \tag{12}$$

In order to evaluate (12), we first use the product decomposition of the involved functions:

$$u(\boldsymbol{x},0) = u_x(x,0)u_y(y,0) \tag{13}$$

with

$$u_x(x,0) = \frac{1}{|I_0|}\chi_{I_0}(x) \tag{14}$$

$$u_y(y,0) = \frac{1}{|J_0|}\chi_{J_0}(y) \tag{15}$$

and

$$\delta(\boldsymbol{x} - \Phi(\boldsymbol{x}')) = \delta(x - \Phi_x(\boldsymbol{x}'))\delta(y - \Phi_y(\boldsymbol{x}')) \,, \tag{16}$$

where the intervals $|I_0|, |J_0|$ are the projections of R_0 onto x- and y-axes, respectively. Correspondingly, Φ_x and Φ_y are the projections of Φ onto x- and y-axes, respectively. These are obtained from (1) as

$$\Phi_x(\boldsymbol{x}') = a_x^\nu + \lambda_x^\nu x' \tag{17}$$

$$\Phi_y(\boldsymbol{x}') = a_y^\nu + \lambda_y^\nu y' \,. \tag{18}$$

Using this factorization, the Frobenius-Perron equation (12) separates into

$$u_x(x,t+1) = \int_{[0,1]} \delta(x - a_x^\nu - \lambda_x^\nu x')u_x(x',t)\,\mathrm{d}x' \tag{19}$$

$$u_y(y,t+1) = \int_{[0,1]} \delta(y - a_y^\nu - \lambda_y^\nu y')u_y(y',t)\,\mathrm{d}y' \tag{20}$$

Next, we evaluate the delta functions according to the well-known lemma

$$\delta(f(x)) = \sum_{l:\text{simple zeros}} |f'(x_l)|^{-1}\delta(x - x_l) \,, \tag{21}$$

where $f'(x_l)$ indicates the first derivative of f in x_l. Eq. (21) yields for the x-axis

$$x_\nu = \frac{x - a_x^\nu}{\lambda_x^\nu} \,, \tag{22}$$

i.e. one zero for each ν-branch, and hence

$$|f'(x'_\nu)| = \lambda_x^\nu \,. \tag{23}$$

Inserting (21), (22) and (23) into (19), gives

$$\begin{aligned} u_x(x,t+1) &= \sum_\nu \int_{[0,1]} \frac{1}{\lambda_x^\nu}\delta\left(x' - \frac{x - a_x^\nu}{\lambda_x^\nu}\right) u_x(x',t)\,\mathrm{d}x' \\ &= \sum_\nu \frac{1}{\lambda_x^\nu} u_x\left(\frac{x - a_x^\nu}{\lambda_x^\nu}, t\right) \end{aligned}$$

Next, we take into account that the distributions must be consistent with the NDA's partition. Therefore, for given $\boldsymbol{x} \in D_\nu$ there is only one branch of Φ contributing a simple zero to the sum above. Hence,

$$u_x(x,t+1) = \sum_\nu \frac{1}{\lambda_x^\nu} u_x\left(\frac{x - a_x^\nu}{\lambda_x^\nu}, t\right) = \frac{1}{\lambda_x^\nu} u_x\left(\frac{x - a_x^\nu}{\lambda_x^\nu}, t\right) \,. \tag{24}$$

Theorem 1 *The evolution of uniform p.d.f.s with rectangular support according to the NDA dynamics Eq. (12) is governed by*

$$u(\boldsymbol{x},t) = \frac{1}{|\Phi^t(R_0)|}\chi_{\Phi^t(R_0)}(\boldsymbol{x})\ . \tag{25}$$

Proof (by means of induction).
1. Inserting the initial uniform density distribution (10) for $t = 0$ into Eq. (24), we obtain by virtue of (14)

$$u_x(x,1) = \frac{1}{\lambda_x^\nu}u_x\left(\frac{x-a_x^\nu}{\lambda_x^\nu},0\right) = \frac{1}{\lambda_x^\nu}\frac{1}{|I_0|}\chi_{I_0}\left(\frac{x-a_x^\nu}{\lambda_x^\nu}\right)\ .$$

Deploying (11) yields

$$\chi_{I_0}\left(\frac{x-a_x^\nu}{\lambda_x^\nu}\right) = \begin{cases} 0 & : \quad \frac{x-a_x^\nu}{\lambda_x^\nu} \notin I_0 \\ 1 & : \quad \frac{x-a_x^\nu}{\lambda_x^\nu} \in I_0\ . \end{cases}$$

Let now $I_0 = [p_0, q_0] \subset [0,1]$ we get

$$\begin{aligned} & \frac{x-a_x^\nu}{\lambda_x^\nu} \in I_0 \\ \iff \quad & p_0 \le \frac{x-a_x^\nu}{\lambda_x^\nu} \le q_0 \\ \iff \quad & \lambda_x^\nu p_0 \le x - a_x^\nu \le \lambda_x^\nu q_0 \\ \iff \quad & a_x^\nu + \lambda_x^\nu p_0 \le x \le a_x^\nu + \lambda_x^\nu q_0 \\ \iff \quad & \Phi_x(p_0) \le x \le \Phi_x(q_0) \\ \iff \quad & x \in \Phi_x(I_0)\ , \end{aligned}$$

Where we made use of (17).

Moreover, we have

$$\lambda_x^\nu |I_0| = \lambda_x^\nu(q_0 - p_0) = q_1 - p_1 = |I_1|$$

with $I_1 = [p_1, q_1] = \Phi_x(I_0)$. Therefore,

$$u_x(x,1) = \frac{1}{|I_1|}\chi_{I_1}(x)\ .$$

The same argumentation applies to the y-axis, such that we eventually obtain

$$u(\boldsymbol{x},1) = \frac{1}{|R_1|}\chi_{R_1}(\boldsymbol{x})\ , \tag{26}$$

with $R_1 = \Phi\ (R_0)$ the image of the initial rectangle $R_0 \subset X$. Thus, the image of a uniform density function with rectangular support is a uniform density function with rectangular support again.

2. Assume (25) is valid for some $t \in \mathbb{N}$. Then it is obvious that (25) also holds for $t+1$ by inserting the x-projection of (25) into (24) using (14), again. Then, the same calculation as under 1. applies when every occurrence of 0 is replaced by t and every occurrence of 1 is replaced by $t+1$.

By means of this construction we have implemented an NDA by a dynamically evolving field. Therefore, we call this representation *dynamic field automaton (DFA)*.

3.2 Kernel Construction

The Frobenius-Perron equation (12) can be regarded as a time-discretized Amari dynamic neural field equation [1] which is generally written as

$$\tau\frac{\partial u(\boldsymbol{x},t)}{\partial t} + u(\boldsymbol{x},t) = \int_X w(\boldsymbol{x},\boldsymbol{x}')f(u(\boldsymbol{x}',t))\ \mathrm{d}\boldsymbol{x}'\ . \tag{27}$$

Here, τ is the characteristic time constant of activation decay, $w(\boldsymbol{x},\boldsymbol{x}')$ denotes the synaptic weight kernel, describing the connectivity between sites $\boldsymbol{x},\boldsymbol{x}' \in X$ and f is a typically sigmoidal activation function for converting membrane potential $u(\boldsymbol{x},t)$ into spike rate $f(u(\boldsymbol{x},t))$.

Discretizing time according to Euler's rule with increment $\Delta t = \tau$ yields

$$\begin{aligned} \tau\frac{u(\boldsymbol{x},t+\tau)-u(\boldsymbol{x},t)}{\tau} + u(\boldsymbol{x},t) &= \int_X w(\boldsymbol{x},\boldsymbol{x}')f(u(\boldsymbol{x}',t))\ \mathrm{d}\boldsymbol{x}' \\ u(\boldsymbol{x},t+\tau) &= \int_X w(\boldsymbol{x},\boldsymbol{x}')f(u(\boldsymbol{x}',t))\ \mathrm{d}\boldsymbol{x}'\ . \end{aligned}$$

For $\tau = 1$ and $f(u) = u$ the Amari equation becomes the Frobenius-Perron equation (12) when we set

$$w(\boldsymbol{x},\boldsymbol{x}') = \delta(\boldsymbol{x} - \Phi(\boldsymbol{x}'))\ . \tag{28}$$

This is the general solution of the kernel construction problem [12, 25]. Note that Φ is not injective, i.e. for fixed x the kernel is a sum of delta functions coding the influence from different parts of the space $X = [0,1]^2$. Note further that higher-order discretization methods of explicit or implicit type such as the Runge-Kutta scheme could be applied to Eq. (27) as well. But in this case the relationship between the Turing dynamics as expressed by the Frobenius-Perron equation (9) and the neural field dynamics would become much more involved. We leave this as an interesting question for further research.

4 DISCUSSION

In this communication we combined nonlinear dynamical automata as implementations of Turing machines by nonlinear dynamical systems with dynamic field theory, where computations are characterized as evolution in function spaces over abstract feature spaces. Choosing the unit square of NDAs as feature space we demonstrated that Turing computation becomes represented as dynamics in the space of uniform probability density functions with rectangular support.

The suggested framework of dynamic field automata may exhibit several advantages. First of all, massively parallel computation could become possible by extending the space of admissible p.d.f.s. By allowing either for supports that overlap the partition of the underlying NDA or for multimodal distribution functions, one could prepare as many symbolic representations one wants and process them in parallel by the DFA. Moreover, DFAs could be easily integrated into wider dynamic field architectures for object recognition or movement preparation. They could be programmed for problem-solving, logical interferences or syntactic language processing. In particular, Bayesian inference or the processing of stochastic grammars could be implemented by means of appropriate p.d.f.s.

For those applications, DFAs should be embedded into time-continuous dynamics. This involves the construction of more complicated kernels through solving inverse problems along the lines of Potthast et al. [12, 25]. We shall leave these questions for future research.

The construction of DFAs has also interesting philosophical implications. One of the long-standing problems in philosophy of science was the precise relationship between point mechanics, statistical mechanics and thermodynamics in theoretical physics: Is thermodynamics merely reducible to point mechanics via statistical mechanics? Or are thermodynamic properties such as temperature emergent on mechanical descriptions?

Due to the accurate analysis of Bishop and Atmanspacher [5], point mechanics and statistical mechanics simply provide two different levels of description: On one hand, point mechanics deals with the dynamics of microstates in phase space. On the other hand, statistical mechanics, in the formulation of Koopman et al. [15, 16] (see Sec. 3), deals with the evolution of probability distributions over phase space, namely macrostates, in abstract function spaces. Both are completely disparate descriptions, none reducible to the other. However, the huge space of (largely unphysical) macrostates must be restricted to a subspace of physically meaningful thermal equilibrium states that obey a particular stability criterium (essentially the maximum-entropy principle). This restriction of states bears upon a contingent context, and in this sense, thermodynamic properties have been called *contextually emergent* by [5].

Our construction of DFAs exhibits an interesting analogy to the relationship between mechanical micro- and thermal macrostates: Starting from microscopic nonlinear dynamics of an NDA, we used the Frobenius-Perron equation for probability density functions in order to derive an evolution law of macrostates: The time-discretized Amari equation (27) with kernel (28). However, with respect to the underlying NDA, not every p.d.f. can be interpreted as a symbolic representation of a Turing machine configuration. Therefore, we had to restrict the space of all possible p.d.f.s, by taking only uniform p.d.f.s with rectangular support into account. For those macrostates we were able to prove that the resulting DFA implements the original Turing machine. In this sense, the restriction to uniform p.d.f.s with rectangular support introduces a contingent context from which symbolic computation emerges. (Note that uniform p.d.f.s also have maximal entropy).

ACKNOWLEDGEMENTS

This research was supported by a Heisenberg grant (GR 3711/1-1) of the German Research Foundation (DFG) awarded to PbG. Preliminary results have been presented at a special session "Cognitive Architectures in Dynamical Field Theory", that was partially funded by an EuCogIII grant, at the 2nd International Conference on Neural Field Theory, hosted by the University of Reading (UK). We thank Yulia Sandamirskaya, Slawomir Nasuto and Gregor Schöner for inspiring discussions.

References

[1] S.-I. Amari, 'Dynamics of pattern formation in lateral-inhibition type neural fields', *Biological Cybernetics*, **27**, 77 – 87, (1977).
[2] J. R. Anderson, *Cognitive Psychology and its Implications*, W. H. Freeman and Company, New York (NY), 1995.
[3] H. Atmanspacher and P. beim Graben, 'Contextual emergence of mental states from neurodynamics', *Chaos and Complexity Letters*, **2**(2/3), 151 – 168, (2007).
[4] H. Atmanspacher and P. beim Graben, 'Contextual emergence', *Scholarpedia*, **4**(3), 7997, (2009).
[5] R. C. Bishop and H. Atmanspacher, 'Contextual emergence in the description of properties', *Foundations of Physics*, **36**(12), 1753 – 1777, (2006).
[6] P. Cvitanović, G. H. Gunaratne, and I. Procaccia, 'Topological and metric properties of Hénon-type strange attractors', *Physical Reviews A*, **38**(3), 1503 – 1520, (1988).
[7] W. Erlhagen and G. Schöner, 'Dynamic field theory of movement preparation', *Psychological Review*, **109**(3), 545 – 572, (2002).
[8] J. Fodor and Z. W. Pylyshyn, 'Connectionism and cognitive architecture: A critical analysis', *Cognition*, **28**, 3 – 71, (1988).
[9] P. beim Graben, 'Incompatible implementations of physical symbol systems', *Mind and Matter*, **2**(2), 29 – 51, (2004).
[10] P. beim Graben, S. Gerth, and S. Vasishth, 'Towards dynamical system models of language-related brain potentials', *Cognitive Neurodynamics*, **2**(3), 229 – 255, (2008).
[11] P. beim Graben, B. Jurish, D. Saddy, and S. Frisch, 'Language processing by dynamical systems', *International Journal of Bifurcation and Chaos*, **14**(2), 599 – 621, (2004).
[12] P. beim Graben and R. Potthast, 'Inverse problems in dynamic cognitive modeling', *Chaos*, **19**(1), 015103, (2009).
[13] J. E. Hopcroft and J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation*, Addison–Wesley, Menlo Park, California, 1979.
[14] M. B. Kennel and M. Buhl, 'Estimating good discrete partitions from observed data: Symbolic false nearest neighbors', *Physical Review Letters*, **91**(8), 084102, (2003).
[15] B. O. Koopman, 'Hamiltonian systems and transformations in Hilbert space', *Proceedings of the National Academy of Sciences of the U.S.A.*, **17**, 315 – 318, (1931).
[16] B. O. Koopman and J. von Neumann, 'Dynamical systems of continuous spectra', *Proceedings of the National Academy of Sciences of the U.S.A.*, **18**, 255 – 262, (1932).
[17] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding*, Cambridge University Press, Cambridge (UK), 1995.
[18] B. McMillan, 'The basic theorems of information theory', **24**, 196 – 219, (1953).
[19] C. Moore, 'Unpredictability and undecidability in dynamical systems', *Physical Review Letters*, **64**(20), 2354 – 2357, (1990).
[20] C. Moore, 'Generalized shifts: unpredictability and undecidability in dynamical systems', *Nonlinearity*, **4**, 199 – 230, (1991).
[21] C. Moore, 'Dynamical recognizers: real-time language recognition by analog computers', *Theoretical Computer Science*, **201**, 99 – 136, (1998).
[22] E. Ott, *Chaos in Dynamical Systems*, Cambridge University Press, New York, 1993.
[23] J. B. Pollack, 'The induction of dynamical recognizers', *Machine Learning*, **7**, 227 – 252, (1991). Also published in [24], pp. 283 – 312.
[24] *Mind as Motion: Explorations in the Dynamics of Cognition*, eds., R. F. Port and T. van Gelder, MIT Press, Cambridge (MA), 1995.
[25] R. Potthast and P. beim Graben, 'Inverse problems in neural field theory', *SIAM Jounal on Applied Dynamical Systems*, **8**(4), 1405 – 1433, (2009).
[26] G. Schöner and E. Dineva, 'Dynamic instabilities as mechanisms for emergence', *Developmental Science*, **10**(1), 69 – 74, (2007).
[27] H. T. Siegelmann, 'Computation beyond the Turing limit', *Science*, **268**(5210), 545 – 548, (1995).
[28] H. T. Siegelmann, 'The simple dynamics of super Turing theories', *Theoretical Computer Science*, **168**, 461 – 472, (1996).
[29] H. T. Siegelmann and E. D. Sontag, 'On the computational power of neural nets', *Journal of Computer and System Sciences*, **50**(1), 132 – 150, (1995).
[30] P. Smolensky, 'On the proper treatment of connectionism', *Behavioral and Brain Sciences*, **11**(1), 1 – 74, (1988).
[31] W. Tabor, 'Fractal encoding of context-free grammars in connectionist networks', *Expert Systems: The International Journal of Knowledge Engineering and Neural Networks*, **17**(1), 41 – 56, (2000).
[32] W. Tabor, 'A dynamical systems perspective on the relationship between symbolic and non-symbolic computation', *Cognitive Neurodynamics*, **3**(4), 415 – 427, (2009).
[33] A. M. Turing, 'On computable numbers, with an application to the *Entscheidungsproblem*', *Proceedings of the London Mathematical Society*, **2**(42), 230 – 265, (1937).
[34] H. R. Wilson and J. D. Cowan, 'A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue', *Kybernetik*, **13**, 55 – 80, (1973).

Machines, life and cognition: a second-order cybernetic approach

Mario Villalobos[1]

Abstract. Building on Maturana's second order cybernetics, in this work I claim that (i) every living being is an autopoietic physical machine, and (ii) every living being is a cognitive system. I contend that what really matters for understanding the cognitive behaviour of living beings is not the fact that they are autopoietic (self-producing) systems, but the fact that they are machines. To recognize that living beings are a particular version of natural machines is to recognize that they are structurally determined systems whose cognitive behaviour works under strict conditions of operational closure. I claim that, from a cognitive point of view, the main consequences of this cybernetic ontology are (1) that, because of their operational closure, living beings cannot exchange information with the outside, (2) that, without receiving information from the outside, they cannot build any internal representation about the environment, (3) that, as natural machines, they are non-teleological systems (i.e. their behaviour is not oriented to any goal), and (4) that, because of their deterministic ontology (i.e. having no possibilities of action or degrees of freedom) they cannot exert any control or regulation over their behaviour. Toward the end I offer a brief comment about the way in which systems such as animats should be conceptualized, avoiding unnecessary philosophical puzzles.

1 MACHINES, MACHINES... AND MORE MACHINES

Cybernetics is, essentially, a formal discipline dedicated to the study of machines in general [1], whatever their constitution (material or formal, real or ideal) and their origin (natural or artificial). From a cybernetic point of view, machines are simply state-determined systems, or, what is the same, systems whose trajectory of changes of state (whose behaviour) follows a deterministic pattern. A system is a machine if its current state, at every moment, is the necessary result of the previous state of the system (when the system is an isolated system), or of the previous state of the system in conjunction with the previous state of its surrounding (when the system interacts with a certain surrounding). For cybernetic purposes the metaphysical status of the system does not really matter. To take an extreme example, entities such as angels or ghosts may be perfectly considered a particular kind of machines. All that is needed is that they behave in a machine-like way [1].

In general, we can distinguish four kinds of machines: 1) Artificial ideal machines; abstract or formal systems created by humans (e.g., a set of algebraic transformations, a syllogism, a Turing machine, etc.). 2) Artificial material machines; man-made physical systems (e.g., a mobile phone, a Watt governor, a car, etc.). 3) Natural material machines; non man-made physical systems (e.g., planetary systems, atmospheric systems, ecological systems, biological systems, etc.). 4) (Super)-Natural ideal machines (conceding a supernatural metaphysics); non man-made immaterial systems (e.g., deities, angels, ghosts, etc.).

It is important to emphasize that the cybernetic concept of machine is highly abstract; it refers to the way in which the system changes its states, and not to the concrete changes of states undergone by the system. What matters is the logic under which certain behaviour is generated, not the behaviour per se. A concrete behaviour may appear to us as more or less rigid or more or less flexible, more or less trivial or more or less intelligent, but that does not matter. Cybernetics is not interested in this kind of description; what matters is the underlying logic that supports such behaviour. For instance, to take an example that we will review in detail later: a drifting boat and a dolphin exhibit very different behaviours. The boat, without helmsman, is dragged by the stream and crashes into the rocks. The dolphin, instead, gracefully navigates avoiding them. The dolphin, but not the boat, exhibits what we call 'intelligent' behaviour. Yet both the boat and the dolphin are equally machines because their respective behaviours, though strikingly different, are sequences of changes of state that follow a deterministic pattern. As we shall see in section 5, the 'special' behaviour of living beings has not to do with whether they are machines or not (living beings are in fact biological machines!) but rather with that they are autopoietic systems. In other words, we will see that the peculiarity of living beings lies uniquely in their organization, not in the way in which they undergo their changes of state.

The same point runs with respect to the relationship between the complexity of certain systems and our epistemic abilities. Something is or not a machine in virtue of a certain behavioural ontology, not in virtue of our epistemic estimations. For example, a pendulum is a simple physical system whose behaviour can be accurately predicted in a relatively easy way (if we know its initial conditions). Why? Because is a deterministic system, a machine. An atmospheric system, on the contrary, is a complex physical system whose behaviour cannot be predicted with the same degree of accuracy; we are forced to use probabilities or statistic estimations. Now, does this mean that the atmospheric system is not a deterministic system, a machine? No, it does not. Every meteorologist knows that atmospheric systems, as any physical system, are systems ruled by natural laws whose changes of state follow a strictly deterministic pattern. An entirely different thing is that we, as observers, are not able to predict (until now) the atmospheric behaviour with absolute accuracy. What the use of probabilities and statistic estimations reveals is not a supposed 'probabilistic' or 'stochastic' ontology in the atmospheric system but rather our own epistemic limitations. The main message here is that the deterministic ontology of

[1] School of Philosophy, Psychology and Language Sciences, University of Edinburgh, EH8 9AD, UK. Email: M.E.Villalobos@sms.ed.ac.uk

machines is independent from our epistemic ability to predict their behaviour.

When we observe any physical system we usually frame it within certain surrounding. We say, for instance, there it is an organism, say, a fungus, in interaction with its environment. When we say that, what do we mean? From a cybernetic point of view, what we mean is that there is a system (the organism) interacting with another system (the environment). What we call environment –that portion of reality outside the system under consideration– is just another machine; a set of variables evolving in a deterministic fashion. The distinction between system and environment is an arbitrary arrangement that, as such, can be changed according to our descriptive purposes. For instance, if I am interested in analyzing the chemical changes undergone by the soil because of the presence of a new colony of fungi, the soil becomes my observed system and the fungi the environment. In the physical realm, both system and environment are always two kinds of machines interacting in a certain way. From the point of view of their behavioural ontologies, the organism and its environment are (to follow with the biological example) strictly symmetric systems. The difference between a system and its environment lies uniquely in their respective organizations and structural compositions, not in the way in which they undergo their changes of state.

So far we have been talking about machines and more machines. The reader, at this point, might wonder whether there is something in the world that could not be considered a machine. Of course, there are things that are not machines, but most of them are ideal entities merely thought by humans. For example, we can think of a system whose behaviour is determined not by its previous states but by its future states. A system whose behaviour is ruled by such teleological logic is not a machine but something different. What we usually call teleological systems are in fact linguistic fictions or ideal artefacts that we use, well or badly, to describe or explain certain behaviours. Nonetheless, teleological systems, as such, do not have physical reality. This point is especially relevant in the study of living beings, whose behaviour is usually interpreted in teleological terms. The risk here is to confuse the teleological character of our descriptive apparatus with the deterministic ontology of the system under description.

For example, classic (first-order) cybernetics considered that teleological descriptions were useful and (almost) indispensable theoretical tools. Nonetheless, it recognized at the same time that the ontology of both artificial and biological systems is always strictly deterministic [2]. Second-order cybernetics, especially in Maturana's version, rejected teleology even as a descriptive tool: living beings are natural machines and any teleological interpretation does nothing more than obscure their deterministic dynamic [3]. So far there is no confusion; both schools understand that biological systems are deterministic systems and that teleology is just an artifice of our descriptions (they just differ with respect to the use of teleological descriptions). The confusion appears only when one conceives teleology as an ontological feature of living beings, as an intrinsic property of their constitution and behavioural dynamic. This is the kind of move made, for example, by approaches like enactive theory. The enactive approach, guided by Jonas's phenomenology of life, contends that living beings are teleological systems; that they are constituted as teleological structures (as natural purposes) and that their behaviour is oriented according to biological purposes [4, 5].

Although in this paper, for reasons of space, we cannot discuss in detail the theoretical relations between the mechanistic ontology of second-order cybernetics and the teleological ontology of the enactive approach, it is quite clear that these approaches work on ontologies that are incompatible. A concrete real system X cannot be, at the same time and in the same respect, a deterministic system (a machine) and a teleological system. That would be equivalent to saying that such system is and is not a machine at the same time. The incompatibility of these ontological positions, as expected, leads to important theoretical and methodological divergences, one of which, as we will see later, has to do with the use of notions such as autonomy, agency or adaptivity. Although the enactive approach and Maturana's second order cybernetics exhibit important affinities (e.g., both of them reject the idea that cognitive systems are information processing systems), it is also clear that their ontological frameworks are, to a large extent, irreconcilable.

2 LIVING BEINGS AS AUTOPOIETIC MACHINES

We have said before that living beings are machines. But, what class of machines are living beings? The second order cybernetics of Maturana is a biological theory that tries to answer this question, focusing on the basic living unity; the cell. Maturana [6] claims that cells, as basic living unities, are autopoietic machines materially realized in the molecular domain, and that every living being is either a cell (a unicellular organism) or an aggregate of them (a multicellular organism).

Living beings are basically a subclass of poietic machines. A poietic machine is a system of production; a network of productive processes. We can distinguish two subtypes of productive machines: allopoietic and autopoietic machines. Allopoietic machines are machines that produce something distinct from themselves (e.g., a car factory), while autopoietic machines are machines that produce themselves. Cells, affirms Maturana, are self-producing or autopoietic machines materially realized in the molecular domain. In other words, cellular systems are molecular networks that produce the molecules that constitute them as such networks. They are organized as a set of relations of production that produce the same components that constitute the system as such [6].

The word 'autopoiesis' denotes a particular kind of organization, not a particular physical reality. Organization is a formal notion, an abstraction. It refers to the set of relations that define the class identity of a system, not to the concrete conditions under which such relations are satisfied or conserved. This means, first, that we have to understand living beings as the molecular version of autopoietic organization, as we could always conceive ideal or formal autopoietic machines. Second, it means that the characterization of autopoietic organization is independent from the conditions under which that organization is conserved in a concrete system. The identification of the organization of a system, whatever it may be, and the identification of the conditions under which that organization is conserved are, strictly speaking, two aspects which are logically independent. I mention this point because some authors tend to misread the notion of autopoiesis introducing in it material and thermodynamic considerations about the viability of a concrete autopoietic system. Material and

thermodynamic considerations are pertinent when we are dealing with concrete cellular systems and their metabolic trajectory, not when we are talking about the autopoietic organization as such.

In the concept of autopoiesis the suffix 'poiesis' is used in its original Greek sense, meaning 'to make', 'to fabricate', or 'to build'. More specifically, the notion alludes to a process of 'synthesis' or 'composition' whereby a set of elements are assembled (combined under certain organization) to form a complex whole. Maturana wants to capture the permanent dynamic of molecular synthesis (formation of molecular compounds, generally organic polymers, by means of one or more chemical reactions) that takes place in the cell metabolism. In this sense, if a car factory is an allopoietic machine, a cell is an autopoietic machine as long as it is a molecular factory that fabricates (synthesizes) the molecules that constitute it as such.

This notion of 'production' as 'synthesis' must be differentiated from the notion of production used in certain philosophical theories of causation, where it is said that an event E has as a cause the event C if E is the effect produced by C (C causes E if C produces E).[2] In this case 'to produce' means simply to bring about or generate a certain state of affairs. For instance, we may say that the increase in temperature produces (causes) the melting of snow, that the friction of bodies produces heat, or that earthquakes produce structural damage in bridges. None of these causal relations, however, involves the assembling of parts or elements to build a complex whole, which is the sense in which 'poiesis' means 'production' (synthesis, composition) in autopoietic theory (see also below the example of the burning candle).

To be a poietic machine in general (allo or autopoietic) a system X has to produce or fabricate something. If X merely renovates components or maintains a certain dynamic without synthesizing anything, then X is not a poietic machine. There are several physical systems that keep constant their organization through a permanent renovation of their material components. For example, a turbulence in the current of a river or a tornado are natural systems that remain constant in their configuration through the renovation of their material components. A cellular system is a system that renovates its material components too, of course, (it needs nutrients and it evacuates chemical waste), and that consumes and dissipates energy (it is a dissipative system), but that is not what defines its class identity as a living system. The difference is that the cell is organized as a productive network, not just as a system through which some components come and go. A tornado or a turbulence, to be an autopoietic system, should be constituted as a network of productive processes, as a factory; they should assemble elements and build the very compounds that constitute them as systems.

There are also systems capable of maintaining their own dynamic more or less stably (within certain parameters) in far-from-equilibrium conditions. A self-sustaining system like a burning candle, for example, is a physicochemical chain of events that maintains itself: the heat of the flame melts and liquefies the wax (combustible solid), this liquefied combustible ascends through the wick (by capillarity) where it is vaporized to finally burn in the flame, whose heat melts and liquefies the wax..., and so on and so forth. At this point the reader, noting the evident circularity of the process, could describe the sequence in causal terms by saying "the heat of the flame causes or produces the melting of the wax, the capillary action produces the ascent of the combustible liquid [...] the combustion produces the flame, whose heat causes or produces the melting...", and then ask "Is not this a productive circular network, and therefore a self-producing or autopoietic system?"

The confusion vanishes if we keep in mind the distinction between the 'poietic' notion of production and the 'causal' notion of production mentioned before. While it is true that the burning candle constitutes a causal circle, it is not the case that this circle is an assembling network that synthesizes the material compounds that constitute the burning candle as such. No new molecular compound is synthesized when the wax melts, the liquid ascends through the wick and vaporizes, or the kinetic energy of gaseous molecules increase; among other things because these processes are either simple state transitions (solid-liquid-gaseous), mere displacements (ascent through the wick), or increases in the molecular kinetic energy (heat).

Taking into account the physical reality of cellular systems and their autopoietic organization, we can say that living beings are a special subclass of dissipative systems; they are poietic systems, and more specifically, systems in which the poietic chain closes on itself.

3 THE KNOW-HOW OF LIVING

Maturana [8] contends that every living being is a cognitive system, and that the natural praxis of living is a cognitive process. Why? Why might every living system be considered a cognitive system? What kind of knowledge or cognition are we talking about? When Maturana [8] says that living beings are cognitive systems, he is making reference to the basic biological 'know-how' that every organism exhibits in the continuous conservation of its autopoiesis, i.e., he is talking about a practical or behavioural knowledge, not about a declarative one. For example, when a protozoan is engulfing a bacterium, it is performing an action, it is doing something, and if the protozoan effectively ends by engulfing the bacterium we cannot but admit that it knows how to engulf bacteria.

We may say, following this reasoning, that every living being, in its continuous doing, reveals a certain know-how that is congruent with its particular form of existence and survival. The protozoan, in doing what it does, reveals the know-how that is characteristic of the protozoic life. The bee, in doing what it does, reveals the know-how that is characteristic of the apiarian life. And every time we see a living being behaving in the way that is characteristic to its particular form of survival, we can say legitimately that such organism constitutes a cognitive system because it exhibits the practical knowledge that corresponds to its own domain of existence. In Maturana's formula: living systems are cognitive systems, and living as a process is a process of cognition [8].

4 COGNITION WITHOUT INFORMATION OR REPRESENTATION

We have defined before, following Maturana, living beings as autopoietic physical machines, and we have also said that cognition corresponds to the continuous doing of living beings. Now we have to think the following: if cognition is the doing of living beings, and if living beings are autopoietic physical machines,

[2] See for example Hall and his "Two concepts of causation" [7].

then cognition is nothing more than the behaviour of such physical machines. That is, nothing more than a particular deterministic sequence of changes of state.

When we address the phenomenon of cognition from a biological point of view, one of the major dangers is to forget that living beings are, despite the amazing plasticity exhibited in their behaviours, nothing more than a subclass of natural machines. We tend to forget that the peculiarity of living beings lies uniquely in their organization (the autopoietic organization), not in the structural logic that generates their behaviour. The way in which a living being undergoes its structural changes is indistinguishable from the way in which a volcano, a boat or a planetary system undergoes its structural changes. A volcano, a boat, a planetary system and a living being exhibit very different behaviours because they are, actually, very different systems. They have different organization and structure. Nonetheless, from the point of view of the structural logic that generates their respective behaviours, volcanoes, boats, planetary systems and living beings are exactly the same; physical machines.[3]

To recognize that living beings are physical machines is to recognize that they are structurally determined systems whose internal functioning takes place under strict conditions of operational closure. In what follows I will argue that, from a cognitive point of view, the main consequences of this ontology are (1) that, because of their operational closure, living beings cannot exchange information with the outside, (2) that, without receiving information from the outside, they cannot build any internal representation about the environment, (3) that, as natural machines, living beings are non-teleological systems (i.e. their behaviour is not oriented to any goal), and (4) that, because of their deterministic ontology (i.e. having no possibilities of action or degrees of freedom for making choices) living beings cannot exert any control or regulation over their behaviour.

The principle of structural determinism [9] establishes that every structural change that takes place in a system occurs: a) because the structure of the system admits such change (otherwise it could not take place), and b) because given the current structural state both of the system and its environment, such change is the only one possible (every structural change is fully determined by the antecedent structural conditions). The structural changes that occur in the system may be consequences of its own internal dynamic or may be triggered by the action of some external agent in the environment. What matters is that in both cases the system always follows its own structural logic. That means that every time the system receives the action of an external factor, it is the current structural state of the system, and not the nature of the triggering element, that specifies the concrete structural change that takes place in the system [9]. For example, if I press the red button on my mobile it turns on, but if I press again the same button the mobile turns off. The changes of state undergone by the mobile do not depend on my finger but on the structural state in which my finger encounters the mobile at each moment. My finger may trigger, but not specify or instruct, such and such structural change in the system. On the other hand, the fact that is my finger and not other element that which triggers certain structural change in the system is, from the point of view of the mobile, absolutely irrelevant. If the mobile is off, it will turn on whenever something, it does not matter what, interacts with the red button in the proper way. That something may be my finger, a friend's finger, a pencil, a stick, a stone, a screwdriver, etc. To the mobile is all the same; simply a transition from 'off' to 'on', nothing more. Its structural dynamic is absolutely blind to the distinctions that we, as external observers, can make about the different triggering objects in the environment.

Every physical machine, natural or artificial, responds as it responds, reacts as it reacts, and does what it does following always its own structural legality. Every machine, in this sense, cannot but be a structurally autonomous system. That is, a system whose changes of state follow the structural legality of the proper system, and where nothing external to the system can infiltrate such legality. Autonomy, so viewed, is a trivial property of every physical system and not, as it is usually thought, a distinctive characteristic of living beings. Living beings are systems trivially –not exceptionally– autonomous. To miss this point leads to a series of mystifications about living beings. For example, it is usually assumed that the autonomy of living beings entails a sort of 'agency'. Living beings, it is said, are 'active' and not merely 'passive' or 'reactive' entities. Instead of 'suffering' certain changes of state, living beings 'do' different things following their own initiative (impulse, drive, motivation, etc.). The general idea is that the behaviour of living beings is not (and cannot be) the mere product of a set of blind mechanisms. Living beings, it is argued, are agents because they enjoy a kind of 'freedom of action'.[4]

These descriptions are mystifications because, as mentioned before in this section, the way in which living systems undergo their structural changes is indistinguishable from the way in which any physical system, natural or artificial, undergoes its structural changes, and that way is a deterministic way. In such circumstances, each structural state in the system leads to the unique structural state possible at that moment, given the structural conditions both of the system and its environment. Under this regime, as is easy to see, systems do not have action alternatives; for better or worse, indeed they cannot avoid doing what they do. But if living beings in general, plant or animal, vertebrate or not, with or without nervous system, cannot avoid doing what they do, do we still want to call them 'agents' in the sense of 'freedom of action'? We will come back to this point later.

From the operational point of view, to say that living beings are structurally determined systems means that their dynamics constitute a closed network of operations or transitions of state [11]. We know that from the material and energetic point of view living beings are essentially open systems. Nonetheless, from the operational point of view, that is, from the point of view of the logic that rules their changes of state, they are closed or auto-defined systems. As analogy, we can think about a dictionary and try to follow its operational logic as a lexical network. No matter in what point of the network (in what word) we start the navigation, it always will send us to another item within the same network, which in its turn will send us to another item again within the same network, and so on in an infinite 'autoreferential' loop. In a similar way, the operations that take place in the structural dynamic of the living system leads always to other operations equally defined by the proper system: living machines are closed domains of transformations [12].

[3] This point is misunderstood even among theoreticians well versed in autopoiesis. See for example [10].

[4] These are typically the interpretations of Enactivists [13, 14, 15].

For example, if we consider the nervous system from this point of view, what we see is a system of internal correlations of activity whose operations always lead to other states of activity within the same system [12]. What we see is an operationally closed system, which does not mean a system that is unable to interact with its medium. A system like the nervous system is actually in permanent interaction with its medium (i.e., in permanent structural coupling), but this interaction does not entail any referential openness [3, 6]. The environment can trigger such and such structural change in the sensory surfaces of the nervous system, but that change remains always 'referred' to the proper structural logic of the system, not to the nature of the triggering element. Like the mobile in our previous example, the nervous system is operationally blind to the nature of the external triggering elements.

The structural dynamic of the nervous system is in a constant state of change, due to its own internal activity and due to the interaction with the environment, and the same goes for the environment. Nervous system and environment are (recall section 1) two symmetric systems from the point of view of their respective structural logics. The environment triggers certain structural changes in the system and the system triggers certain structural changes in the environment, so their respective structural trajectories remain necessarily tied. In other words, they remain in permanent structural coupling. This structural coupling is, if you will, like a permanent structural dance, in the sense that what one observes is a congruence or coherence between their respective structural changes. This behavioural coherence (that we call cognition) nonetheless, is the result of this historical process of recurrent interactions and not the result of an alleged 'internal representation' of the environment inside the nervous system.

It is a mistake, therefore, to describe the nervous system as having openings or 'windows' to the outside. Even worse is the idea that, through these supposed 'openings', the nervous system receives certain informational contents. In general, between a structurally determined system (alive or not) and its environment there can be no transfer of contents or informational specifications because their respective structural logics, though coupled, remain always independently auto-defined functioning in operational closure.

This observation contradicts the idea that living beings are systems open to information; that they are 'informavore' systems [16]. The informational metaphor presents living beings as systems that consume or ingest information, just as if it were a kind of cognitive food. This information is in the environment and living systems pick it up, consume it and process it.[5] Perception, according to this view, is precisely the mechanism of picking up information from the environment, processing it and interpreting it in a certain way [17]. The general idea is that living beings pick up information in order to build internal representations of the external world.[6] The acquisition ('ingestion') of information is vital because without informational content, without this raw material, the organism cannot elaborate any kind of internal representation (i.e., there is no representation without information).

[5] Authors like Millikan [18], for example, consider that a bad cognition (a wrong belief) is analogous to a bad digestion.

[6] There are also theories that do not posit internal representations but only a direct collection of information. From our point of view these ecological approaches of perception are equally misleading [19, 20].

Well, what we have shown here is precisely that living systems, due to their operational closure and structural determinism, are unable to 'ingest' informational contents. To follow with the digestive metaphor, what we have shown is that living beings are 'information intolerant'. But if living beings cannot ingest information, if they cannot get this raw material, then they cannot build internal representations either.

Living systems in general, and nervous system in particular, do not work on the basis of internal representations. Strictly speaking, taking into account the structural and operational ontology of living beings, the so called 'internal representations' have no biological reality.

5 INTELLIGENT BEHAVIOR WITHOUT CONTROL OR PURPOSE

When we see the plastic and intelligent behaviour of living beings, we tend to forget that we are dealing with natural machines. That is, with systems wherein every structural state is the necessary outcome of the previous structural conditions (both of the system and of the environment). We forget that, under a regime like this, living beings have neither action possibilities nor degrees of freedom for making choices.

On the contrary, if we remind that living beings are autopoietic machines, we realize that they cannot help but react and act the way they do (just like the rain cannot but go from the sky to the ground). We understand that, having no action possibilities, living beings cannot exert any control or regulation over the trajectory of their structural changes [9].

In a similar way, we tend to forget that living beings, like any natural physical system (volcanoes, planetary systems, etc.), are purposeless systems. Since every change of state is fully determined by its respective antecedent structural state, there is no any future state, purpose, end or goal in general that can have any effective participation in the realization of the behaviours of the system [9].

Certainly, here we are facing the metaphors of living beings as control systems (i.e., self-regulatory systems) and teleological systems (i.e., systems oriented at a goal). These metaphors were a key element in the first generation of cybernetics (1940-1960), when Wiener canonically defined cybernetics as the science of control and communication [21], and animal behaviour was basically understood as teleological behaviour [2]. The organism, and specially the nervous system, was viewed as a kind of 'pilot' or 'helmsman' that guided the actions; monitoring, controlling and correcting the movements according to certain goals.

Maturana, in order to avoid this kind of metaphor, offers the notion of 'structural drift' [9]. Living beings, says Maturana, are like drifting boats. Without helmsman, without a navigation plan, they simply follow a trajectory of structural coupling with the environment. Living beings neither control nor regulate their movements. They neither anticipate nor foresee future results because their behaviour, being a mechanistic process of structural coupling, is not oriented to any goal or biological purpose [22]. The notion of structural drift may seem counterintuitive only if we misread the analogy. Maturana does not contend that a drifting boat and, say, a dolphin, behave in the same way. What he contends is that, although boat and dolphin behave in a very different way, the underlying structural logic that generates their respective behaviours is the same. Dolphin behaviour is peculiar because its structural organization is peculiar, full stop.

The dolphin, not the boat, is constantly producing itself. The structural coupling that he establishes with the environment generates, naturally, a trajectory of structural changes that are characteristic of autopoietic systems; structural changes that the boat cannot undergo because its structural organization is completely different.

The old metaphor of living beings as control systems oriented at a goal, firmly rejected by Maturana in his second-order cybernetics, is still appealing to many people. The typical assumption here is that the intelligent and flexible behaviour of living beings is the result of a set of internal processes that control, regulate, monitor and guide the behaviour. The nervous system is viewed as a communicative control system (more or less homuncular) [23] that regulates and facilitates the actions of the organism through a set of instructions and informational messages. Usually, though not always, these informational neural mechanisms are also viewed as internal (context-dependent, weak, minimal, teleonomic) representations oriented to action [18, 24, 25]. The intelligent behaviour of living beings is considered 'intelligent' precisely insofar as it is mediated by these supposed internal motor-representations [26, 27].

Nevertheless, the assumption of internal representations is not a necessary condition for embracing this old fashioned first-order cybernetic metaphor. Enactivist philosophers and scientists, for example, reject the idea that the intelligent action of living beings is a phenomenon mediated by internal representations [28, 29]. Instead, they think that living beings are special because of their 'adaptivity' [5, 15]. According to enactivists, and in a jargon that recalls the golden age of classic cybernetics, adaptivity is a distinctive property of living beings that has to do essentially with processes of self-monitoring, self-regulation and control (of both internal dynamics and external exchanges) [5]. Moreover, this time inspired by Merleau-Ponty's phenomenology, enactivists, but also Heideggerians like Dreyfus [30], defend the idea that the action of living beings exhibits a peculiar kind of intentionality, one that is more fundamental (and universal) than referential (aboutness) and content (semantic) intentionality. According to these authors, actions are intentional in the sense that they are directed at some goal or purpose; that they work as a means to a further end [31, 32, 33]. This is, as we have seen before, basically a teleological interpretation of behaviour, and, by extension, of living systems in general. Living beings are viewed as systems endowed with both intrinsic and projective teleology [5, 14].

With or without internal representations, it seems the classic cybernetic metaphor of living beings as control systems oriented at a goal is still, unfortunately, quite attractive. This metaphor, as we have seen, is doubly misleading. On the one hand, it creates the illusion that living beings, managing action possibilities, select their behaviour according to certain internal (biological) norms. And on the other hand, it obscures the deterministic ontology of living beings as autopoietic machines, depicting them as teleological systems.

6 FINAL COMMENTS

In this paper we have said that living beings are autopoietic physical machines, and that their natural behaviour as biological systems is a cognitive process. We have underlined the cybernetic ontology of living beings as natural machines, trying to show that biological phenomena are, despite their peculiarity, strictly continuous with the rest of natural phenomena.

This continuity, from the conceptual or terminological point of view, entails, for example, that expressions like 'bio-machine hybrids' and similar are not completely correct. From a cybernetic perspective, the expression 'bio-machine', or 'biological machine' in general, does not denote any ontologically hybrid condition because, after all, biological machines is what living beings actually are! The expression 'bio-machine' is, at most, another way to say 'living being' (or 'autopoietic physical machine'). The design and construction of systems that conjugate living and nonliving components, like animats, are hybrid just in the sense that there are two different kinds of physical machines (machines with different structural organization) assembled into a single system; autopoietic and non autopoietic machines. Beyond this organizational distinction, there is nothing 'deeply hybrid' –as some enactivists might think– in these kinds of systems. For example, animats should not be conceived as systems made of teleological and non-teleological components, intentional and non-intentional elements, autonomous and heteronomous parts, and so on and so forth. This kind of interpretation, generally inspired by phenomenological considerations, can only bring us unnecessary philosophical puzzles.

We have defended the idea that living beings are physical machines, not that they are (necessarily) natural machines. Today we are dealing with animats, but tomorrow we may deal with fully artificial autopoietic machines. Living beings may be, without any contradiction, artificial machines too.

Whatever their origin (natural or artificial), here we have said that what matters from a cognitive point of view is that living beings are structurally determined systems that work under strict conditions of operational closure. We have highlighted the fact that, although living beings exhibit a remarkable behavioural plasticity, such plasticity cannot be attributed to properties or mechanisms that do not have a place within the dynamic of the deterministic systems and their structural drift. We have seen that notions such as control, self-regulation, purpose or goal-oriented behaviour, though familiar to our common sense, do not map on to any real phenomenon in the functioning of living systems [34]. We have also seen that, given that living beings work under conditions of operational closure, the idea that they are open systems that receive or pick up informational contents from the environment is unsustainable. We have stressed the fact that, under such conditions, the structural states and dynamics of the system cannot establish any kind of referential relation with any state of affairs in the environment, and that consequently 'internal representations' are explanatory fictions that have no biological reality.

AKNOWLEDGMENTS

Many thanks to Dave Ward, Mark Sprevak and Robert O'Shaughnessy for their academic support. Also many thanks to the anonymous reviewers for helpful comments on this work.

REFERENCES

[1] Ashby, W. R. (1962). Principles of the self-organizing system. In H. Von Foerster and G. W. Zopf, Jr. (Eds.), *Principles of Self-Organization:Transactions of the University of Illinois Symposium* (pp. 255-278). London, UK: Pergamon Press.

[2] Rosenblueth, A., Wiener, N., & Bigelow, J. (1943). Behaviour, Purpose, Teleology. *Philosophy of Science, 10 (1),* 18-24.

[3] Maturana, H.R. & Varela, F.J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht, Holland: Kluwer Academic Publishers.

[4] Weber, A., & Varela, F. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences 1*, 97–125.

[5] Di Paolo, E. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences, 4 (4),* 429–452.

[6] Maturana, H. (1975). The organization of the living: A theory of the living organization. *International Journal of Man-Machine studies, 7*, 313-332.

[7] Hall, N. (2004). Two Concepts of Causation. In J. Collins, N. Hall and L.A. Paul (Eds.) *Causation and Counterfactuals* (pp. 225-276). Cambridge: MIT Press.

[8] Maturana, H. (1970). Biology of cognition. Reprinted in H. Maturana & F. Varela, (1980) *Autopoiesis and Cognition: The Realization of the Living* (pp. 5-56). Dordrecht, Holland: Kluwer Academic Publishers.

[9] Maturana, H. (1987). Ontology of Observing: The biological foundations of self consciousness and the physical domain of existence. In E. Caianiello (Ed.), *Physics of Cognitive Processes* (pp. 324-379). Singapore: World Scientific.

[10] Barandarian, X., & Moreno, A. (2008). Adaptivity: From metabolism to behaviour. Adaptive behaviour, 16 (5), 325-344.

[11] Bourgine, P., & Varela, F. (1992). Towards a Practice of Autonomous Systems. In F. J. Varela & P. Bourgine (Eds.), *Toward a Practice of Autonomous Systems. Proceedings of the first European conference on artificial life* (pp. xi-xvii). London: MIT.

[12] Maturana, H., & Varela, F. (1987). *The Tree of Knowledge*. Shambhala New Science Library: Boston and London.

[13] Di Paolo, E. A., Rohde, M. & De Jaegher, H. (2010). Horizons for the Enactive Mind: Values, Social Interaction, and Play. In *Enaction: Towards a New Paradigm for Cognitive Science*, ed. J. Stewart, O. Gapenne and E. A. Di Paolo, 33-87. Cambridge, MA: MIT Press.

[14] Thompson, E. (2007). *Mind in Life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: Harvard University Press.

[15] Froese, T., & Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence, 173*, 466-500.

[16] Pylyshyn, Z. (1986). *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge, Mass.: MIT Press.

[17] Marr, D. (1982). *Vision*. San Francisco: WH Freeman.

[18] Millikan, R.G. (1993). *White queen psychology and other essays for Alice*. Cambridge, MA: MIT Press.

[19] Gibson, J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

[20] Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA.: MIT Press.

[21] Wiener, N. (1948). *Cybernetics: or Control and Communication in the Animal and the Machine*. Cambridge, Mass.: MIT Press.

[22] Maturana, H. (2008). Anticipation and Self-consciousness. Are these functions of the brain? *Constructivist Foundations*, *4 (1),* 18-20.

[23] Wheeler, M. (2005). *Reconstructing the Cognitive World*. Cambridge, Mass.: MIT Press.

[24] Clark, A. (1997). *Being there: Putting brain, body and world together again*. Cambridge, Mass.: MIT Press.

[25] Menary, R. (2007). *Cognitive Integration*. New York: Palgrave.

[26] Clark, A., & Grush, R. (1999). Towards a cognitive robotics. *Adaptive Behavior*, *7*, 5-16.

[27] Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, *27*, 377-442.

[28] Gallagher, S. (2008). Are minimal representations still representations? *International Journal of Philosophical Studies*, *16 (3)*, 351-369.

[29] Varela, F., Thompson, E., & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and human experience*. MIT Press: Cambridge.

[30] Dreyfus, H. L. (2008). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. In P. Husbands, O. Holland & M. Wheeler (Eds.) *The Mechanical Mind in History* (pp. 331–71). Cambridge, Mass.: MIT Press.

[31] Gallagher, S., & Miyahara, K. (2011). Neo-pragmatism and enactive intentionality. In J. Schulkin (Ed.) *Action, perception and the brain: Adaptation and cephalic expression*. Basingstoke, UK: Palgrave-Macmillan.

[32] Menary, R. (2009). Intentionality, cognitive integration and the continuity thesis. *Topoi*, *28*, 31-43.

[33] Hutto, D. (2011). Philosophy of mind's new lease on life: Autopoietic enactivism meets teleosemiotics. *Journal of Consciousness Studies*, *18(5-6)*, 44-64.

[34] Maturana, H. (2011). Ultrastability ... Autopoiesis? Reflective response to Tom Froese and John Stewart. *Cybernetics and Human Knowing*, *(18)1-2*,143-152.

Mind and artifact: A multidimensional matrix for exploring cognition-artifact relations

Richard Heersmink

Dept. of Cognitive Science, Macquarie University, Sydney

Abstract. What are the possible varieties of cognition-artifact relations, and which dimensions are relevant for exploring these varieties? This question is answered in two steps. First, three levels of functional and informational integration between human agent and cognitive artifact are distinguished. These levels are based on the degree of interactivity and direction of information flow, and range from monocausal and bicausal relations to continuous reciprocal causation. Second, a multidimensional framework for exploring cognition-artifact relations is sketched. The dimensions in the framework include reliability, durability, trust, procedural and representational transparency, individualization, bandwidth, speed of information flow, distribution of computation, and cognitive and artifactual transformation. Together, these dimensions constitute a multidimensional space in which particular cognition-artifact relations can be located. The higher a cognition-artifact relation scores on these dimensions, the more integration occurs, and the more tightly coupled the overall system is. It is then better, for explanatory reasons, to see agent and artifact as one cognitive system with a distributed informational architecture.

1 INTRODUCTION

There is a great variety in both the kinds of cognitive artifacts and the cognitive profiles of the human agents that use those artifacts. Due to this variety, a multiplicity of relations is established between agents and cognitive artifacts. One way to look at these relations is through the lens of extended mind theory (EMT), according to which some of these relations ought to be seen as constitutive. EMT argues that human cognition is in certain cases constituted by an embodied human brain and cognitive artifacts [1]. Consequently, cognitive artifacts are not seen as merely external aids or scaffolds for thinking, but are sometimes proper and constitutive parts of an extended or distributed cognitive process. Cognitive processes are thus conceptualized as hybrids or amalgamations of neurological, bodily, and environmental processes [2].

John Sutton has identified two movements or waves in EMT. The first wave is mostly based on the "parity principle" and is advocated by Andy Clark and David Chalmers [3], Mike Wheeler [4, 5], and others. The second wave is based on what Sutton calls the "complementarity principle" and is advocated by Sutton [6, 7], Richard Menary [8], Clark [9], and Julian Kiverstein and Mirko Farina [10]. The parity principle reads as follows: "If, as we confront some task, a part of the world functions as a process which, were it done in the head, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world is (so we claim) part of the cognitive process" [3, p. 12]. Thus, according to this principle, a cognitive process is extended when a cognitive artifact (or other part of the world) functions in a *similar* way as a clearly recognized internal cognitive process. So the parity principle invites us to see similarities between internal and external states and processes as a sufficient condition for cognitive extension.

The complementarity principle, in contrast, reads as follows: "In extended cognitive systems, external states and processes need not mimic or replicate the formats, dynamics or functions of inner states and processes. Rather, different components of the overall (enduring or temporary) system can play quite different roles and have different properties while coupling in collective and complementary contributions to flexible thinking and acting" [6, p. 194]. So in contrast to parity, complementarity argues that cognitive artifacts need not have similar functions to internal processes, but often *complement* internal processes with *different* properties and functionalities. In fact, complementing brain functions is often the point of deploying cognitive artifacts: so that they can perform functions the brain cannot do or cannot do well. Jointly, brain-plus-artifact is a much more versatile and powerful problem-solving system than the brain alone.

The parity principle stresses functional isomorphism, and downplays differences between internal and external states and processes, implying that the nature and properties of cognitive artifacts as well as their impact on our brains and behavior do not really matter. It also downplays individual differences between humans and how they interact with cognitive artifacts. There are clear differences in how different humans rely on and deploy cognitive artifacts. Some people remember their appointments, while others rely on their agenda, some people are good at mathematics, while others need pen and paper to solve calculations, et cetera. Such differences ought to be and are conceptually and empirically studied, but the parity principle does not encourage such interdisciplinary study of the variety of relationships between human agents and cognitive artifacts.

Contrary to parity-based EMT, complementarity-based EMT *does* encourage interdisciplinary study of different kinds of interaction and coupling. So rather than providing a new metaphysics of the mind, as Sutton remarks, it encourages detailed case studies in which the differences between the contributing elements are analyzed. The goal of this paper is to further develop complementarity-based EMT by sketching a multidimensional framework for better understanding and exploring the different kinds of epistemic interactions between humans and cognitive artifacts. The dimensions in the framework include reliability, durability, trust, procedural and representational transparency, individualization, bandwidth, speed of information flow, distribution of computation, and cognitive and artifactual transformation. There are at least two reasons why we need such a multidimensional framework. First, because it encourages and provides a conceptual toolbox for a detailed interdisciplinary study of the variety of relationships between human agents and cognitive artifacts. Second, given that a substantial part of our cognitive activity quite heavily depends on cognitive artifacts, it is important to have a framework that gives us a richer and deeper understanding of the interactions with such artifacts. If we are "natural-born cyborgs" (i.e.

inherently tool-using and tool-incorporating creates), then it is important to better understand the variety of relationships that are established between us and our cognitive tools.

Importantly, although motivated by EMT, this framework is not restricted to the extended mind cases in which a minimal requirement is a two-way or reciprocal interaction. Clark and Chalmers [3] have characterized extended minds as "coupled systems" in which there is a *two-way* interaction between an agent and cognitive artifact. Agent and artifact both play an active causal role in an overall cognitive process. Their notion of a coupled system requires two-way interaction between agent and artifact. But, it is still important to better understand one-way or monocausal relations. Indeed, a large amount of cognitive artifacts have a monocausal influence on human thought and behavior, so for explanatory reasons (i.e., explanatory scope and completeness) it would be unwise to exclude monocausal relations from the picture even if these cases are not candidates for extended cognition. In order to develop a more inclusive picture and to better understand the causality of information flow in cognition-artifact systems, I first say a few words on the concept of a cognitive artifact and then distinguish between three levels of information flow, including monocausal and bicausal relations as well as continuous reciprocal causation between human and artifact. Thereafter, I outline the multidimensional framework in which most of the dimensions are also helpful for better understanding monocausal interactions. And I end with briefly applying the framework to two concrete examples.

2 COGNITIVE ARTIFACTS

In an influential paper, Donald Norman has defined a cognitive artifact as "an artificial device designed to maintain, display, or operate upon information in order to serve a representational function and that affect human cognitive performance" [11, p. 17]. A cognitive artifact is thus a device that is intentionally designed to serve a representational function that has an influence on human thought. There is a wealth of such devices in our environment, including road signs, maps, diagrams, notebooks, thermometers, agendas, textbooks, smart phones, tablet computers, PowerPoint presentations, navigation systems, software programs, laptops, and desktops.

Importantly, the representational formats of cognitive artifacts need not always be linguistic, pictorial, numerical, or diagrammatic. For example, when I leave an empty milk bottle on the kitchen dresser to remind myself that I need to buy milk, one could argue that the milk bottle is a mnemonic aid. Although the representational format is not linguistic, pictorial, numerical, or diagrammatic, it still refers to a certain task. Likewise, in order to reduce their memory load, bartenders learn to use the shapes of drink glasses and their placement on a bar as a material representation for the sequence of the ordered drinks [12]. Here a distinction can be made between a *designed* and *improvised* representational function. Milk bottles and drink glasses have not been designed as mnemonic aids, however, during improvisation we can attribute cognitive or representational functions to initially non-cognitive artifacts. An important point here is that a cognitive artifact is not defined by some intrinsic properties, but because of the way it is used.

In some cases, even non-technological objects or structures serve representational functions. For example, some seafarers can navigate with the help of celestial objects such as the sun or stars. Seafarers, in such cases, do not navigate with the help of a cognitive artifact in the sense of a human-made device with a representational function like, for instance, a compass or radar system, but navigate with the help of a naturally occurring structure. The sun or stars, then, have a representational function which is attributed to them by a human agent or group of human agents.

Thus, as Edwin Hutchins informs us, "there is no widespread consensus on how to bound the category of "cognitive artifacts." The prototypical cases seem clear, but the category is surrounded by gray areas consisting of mental and social artifacts, physical patterns that are not objects, and opportunistic practices" [13, p. 127]. Given that we attribute representational functions to initially non-representational artifacts and naturally occurring structures, and perform actions on the basis of the information those artifacts and structures represent, I would like to propose a more liberal definition of cognitive artifacts as to include any object or structure (human-made or naturally occurring) with designed or improvised representational functions that affect human cognitive performance.

3 LEVELS OF INFORMATION FLOW

Having briefly looked at what a cognitive artifact is, let us now look at how information flows between humans and cognitive artifacts. I identify three levels of information flow, which should not be seen as clearly distinguished, but as overlapping. The first level is characterized by a monocausal or one-way information flow from artifact to agent. Examples include clocks, compasses, slide rulers, road signs, maps, dictionaries, encyclopedias, cookbooks, websites, documentaries, no-smoking signs, memorials, graphs, celestial objects, diagrams, manuals, and timetables. Humans make decisions and structure their actions on the basis of the information that such cognitive artifacts provide. We depart to the train station when our watch tells us it is time to go, we take a left turn because the map says it is the shortest route to our destination, and so forth. Further, and this is essential, the agent typically does not have any influence on the content and nature of the information. Such artifacts and the information they carry are designed, installed, and distributed by other human agents including writers, designers, publishers, news agencies, companies, governmental institutions, et cetera. Thus, one can argue that such cognitive artifacts and symbol systems *mediate* information flow between a designer and user.

In some cases, we interact bodily with the artifact in order to obtain the information we need, i.e., it is an interactive process. We interact bodily with compasses, slide rulers, maps, cookbooks, and manuals to get the information we need, in which case there are three steps involved: bodily interaction, perceptual intake, and action. In other cases, we merely have to look at the artifact to extract the relevant information. No-smoking signs, road signs, and memorials, for example, need not be interacted with bodily to obtain the information we want, in which case there are two steps involved: perceptual intake and action. However, not every deployment of a cognitive artifact results in an action. Occasionally, we are inhibited from performing an action. No-smoking signs and no-parking signs, for example, do not result in an action, but in an inhibition of an action (but only for those who would have otherwise smoked or parked, of course).

The second level is characterized by a bicausal or two-way information flow, i.e., from agent to artifact and then from artifact to agent. Humans frequently offload information onto their environment to relieve their memory burdens, in that way creating cognitive artifacts such as post-it notes, notebook and agenda entries, shopping-lists, to-do lists, and lists of addresses, birthdays, and telephone numbers; but also artifacts with improvised representational functions like empty milk bottles and drink glasses. Artifacts in two-way relations are often tailored for individual use and are frequently not part of publicly available artifacts or symbol systems (e.g., road signs, clocks, and textbooks), although there are exceptions such as a shared agenda. They are closed systems in the sense that the cognitive artifact is meant for one individual agent who has designed the informational content of the artifact for individual use. In one-way systems, designers outside the system have designed both the physical structure and informational content of the cognitive artifact. But in two-way systems, designers have designed the physical structure of the artifact, e.g. the structure of a post-it note or agenda that enables a user to offload information onto the artifact. However, and this is essential, the informational content and representational function is designed by the user.

In terms of repeatability, there are different versions of two-way relations. On one side of the spectrum there are one-offs like, for example, a post-it note with a brief reminder such as the date of a deadline. This example is a one-off, as there is only one cycle of offloading, intake, and action. On the other side of the spectrum there are often repeated interactions with a single artifact as in the case of Otto and his notebook [3]. Otto writes important information in his notebook and then consults it to act on the basis of that information. He usually does not further manipulate the existing information in the notebook, but does on occasion add new information to it when he needs to do so. The content of new entries in the notebook usually does not depend on the content of previous entries. When Otto writes down the address of MoMa, it is because he knows that in the future he might be going to MoMa and for some external reason he is triggered to write down its address. It is not because other information in the notebook triggered him to do so. With post-it notes there is in most cases one cycle of offloading, intake, and action. But with Otto's notebook there are various distinct cycles of offloading, intake, and action, which are repeated over a certain period of time. However, the informational content of each cycle does not depend on the informational content of previous ones. Hence, Otto and his notebook form a two-way system, just one that is often repeated.

The third level is based on a reciprocal information flow. Occasionally, cognitive artifacts are integral parts of ongoing information-processing tasks. Writing an academic paper [14], making a PowerPoint presentation, solving a difficult calculation with pen and paper [15], or designing an architectural blueprint of a building often involves small incremental steps. We do not have a finished paper, presentation, calculation, or architectural blueprint in our head and then fully offload it onto the artifact. Rather, we offload small bits of information onto the artifact, and the nature and content of the offloaded information contributes to and partly determines the next step in the overall process. For example, when writing an academic paper one often starts with a rough outline, which may prompt ideas about how to fill in the details. Filling in the details may then prompt an adjustment of the outline, which may in turn prompt further details. This process may continue for a number of cycles. Each step in the overall process builds and depends on previous steps. The human agent and cognitive artifact continuously exchange information and so there is a reciprocal and cumulative information flow that constantly transforms the cognition-artifact system. There is, in Clark's words, "continuous reciprocal causation" between agent and artifact [9].

Like information flow in two-way systems, reciprocal information flow often takes place in a closed system in the sense that the cognitive artifact is meant for a single human agent who has designed the informational content of the artifact for individual use. In two-way relations, there were three steps involved: offloading, intake, and action. This is roughly the same for reciprocal relations, except that each cycle depends on the outcome of the previous one. The cycles are thus *interdependent*. For this reason, the functional and informational integration between agent and artifact is significantly closer than in one-way and two-way systems. It is not a mere exchange of information between two entities, as in two-way systems. What is offloaded onto the artifact in a given cycle depends on what is offloaded in the previous cycle(s) and, therefore, the degree of hybridization and integration is considerably higher. In fact, this integration is so dense that it is better to conceive of agent and artifact as one cognitive and information-processing system.

To be complete, there is a fourth level of information flow (which may be called system information flow) that, for reasons of scope, has not been outlined here. Information quite often flows in systems that are comprised of more than one human agent and more than one cognitive artifact. Examples include a team of engineers working on the design of a car, researchers in a scientific laboratory, and pilots in the cockpit of an airplane. These are cases of collective cognition in the sense that there is a (more or less integrated) collective that tries to solve a particular problem by using cognitive artifacts. Extended cognition and collective cognition in the sense just explained are related but distinct phenomena. The focus in this paper is on *agent*-artifact relations and extended cognition, so I only focus on the first three levels of information flow.

4 A MULTIDIMENSIONAL FRAMEWORK

Having identified these three levels of information flow in cognition-artifact relations, let us now continue with outlining the dimensions that are important for further exploring these levels. Sutton [16, 7], Clark and Robert Wilson [17], and Kim Sterelny [18] have articulated the idea of a dimensional analysis of cognition-artifact relations. Clark and Wilson identify two dimensions: first, the nature of the non-neural resources, which may be natural, technological or socio-cultural; and second, the durability and reliability of the overall system. Sutton [7] takes the dimensions of reliability and durability as well as the dimensions of trust and glue mentioned earlier by Clark and Chalmers [3], and also briefly mentions the dimension of transparency. And finally, Sterelny discusses three dimensions, namely, those of trust, individualization, and individual versus collective use. Their dimensional frameworks are perceptive and insightful, but tend to emphasize certain dimensions while overlooking others.

In this section I aim to refine and synthesize their dimensions into a coherent and systematic multidimensional framework, add a number of dimensions to the framework, and briefly examine where and how some of these dimensions overlap and interact.

Note that this framework is not meant as an exhaustive list; there may be other dimensions relevant for better understanding cognition-artifact relations, and the dimensions are rather sketchy for reasons of space. Before outlining and discussing each dimension, it is helpful to distinguish a number of elements that are relevant for better understanding the underlying conceptual structure of each dimension. These elements are: (1) the cognitive profile or cognitive capacities of the human agent; (2) the representational, functional, and technical properties of the cognitive artifact; (3) the task environment and context of use; and (4) the kind of epistemic action and its epistemic purpose. Although essential for better understanding cognition-artifact coupling, these elements are not dimensions, but each dimension emerges out of the interplay between two or more of these elements. These dimensions are all matters of degree and relational in the sense that they never depend on only one of those elements.

To give a brief example: the dimension of trust emerges out of a specific epistemic interaction between agent and artifact performed in a particular context and with a specific epistemic purpose in mind. Some artifacts like an authoritative textbook on some subject (say, astrophysics) are almost automatically trusted, while others like Wikipedia are trusted with much more care. So trust depends on the properties of the artifact, but also on one's cognitive attitude towards the artifact, which may differ from agent to agent. Some people may trust Wikipedia by default, while others are highly skeptical of its truth-value. Our cognitive attitude towards information also depends on the context. Libraries and universities, for example, are generally seen as contexts in which trustworthy information can be found. But the information provided by the ministry of truth in George Orwell's dystopian novel, *1984*, is likely to be encountered with skepticism. The dystopian world that Orwell describes is a context in which information is distrusted because it is provided by a government that constantly gives misinformation. So whether a human agent trusts certain information depends on a number of elements, namely: the cognitive profile of the agent, properties of the artifact, context, and the purpose of the epistemic action. Let us now turn to the dimensions.

4.1 RELIABLE ACCESS

Reliable access to external information is one of the key dimensions for how and how often an epistemic interaction unfolds [3, 17]. Several things are important here. First, the cognitive profile of the agent partly determines the necessity for information access. Some people have bad memory capacities and therefore rely and depend more on memory aids such as post-it notes, agendas, and other reminders. Other people have bad mathematical skills and rely and depend more on calculators or perform calculations with pen and paper. While yet other people have bad navigation skills and rely and depend on navigation aids such as road signs, maps, and navigation systems. There are also people who have better memory, mathematical, or navigational skills and do not or rely less on external artifacts.

Second, reliability depends on the kind and properties of the artifact. Due to their technical properties, some artifacts provide better information access than others. Take agendas, for instance. As long as one does not forget to bring one's analogue agenda when needed, it provides reliable access to the information in it. In contrast, digital agendas embedded in one's smart phone, tablet, or other electronic device, in one sense, provide less reliable access, because they are inaccessible without electricity. So one not only needs to remember to bring the device when needed, but also to charge it when the battery is empty. Further, digital cognitive artifacts can potentially malfunction in more ways than analogue ones. So next to battery issues, there may be numerous software and hardware issues that prevent one from accessing one's digital agenda. Software and hardware issues are irrelevant for analogue agendas. But, conversely, digital agendas such as Google Calendar are online systems that store information in the cloud and are therefore less susceptible for theft or loss than analogue agendas. Even if one looses one's wearable computing device, the information is still available in the cloud. Analogue agendas lack these properties.

Third, the context and kind of epistemic action are relevant for reliable information access. A carpenter only brings his slide ruler when he effectively needs it, which is during work. Carpenters only need access to slide rulers when they need to perform the epistemic action of measuring the length of some object. Such epistemic actions are frequently performed during work and thus in a work-environment. Carpenters presumably do not bring slide rulers to the supermarket or dinner parties, although there may be exceptions, because there is nothing for them to measure (unless they are working in a supermarket or at dinner parties, of course). So necessity of information access depends on the kind of epistemic action and context. Certain epistemic actions are thus only performed in particular contexts.

4.2 DURABILITY

There are two sides to durability: first, the durability of the artifact itself, and second, the durability of the relationship with the artifact. Certain cognitive artifacts are highly durable, while others are less durable. When handled carefully - textbooks, abacuses, and slide rulers can potentially last for decades, whereas analogue agendas last for roughly a year, and shopping-lists and to-do lists often last for just a few hours. This depends on both the material quality and properties of the artifact as well as the purpose of the epistemic action. Generally, the better the material quality of the artifact, the more durable it is.

But, more importantly, the durability and repeatability of our relationship with cognitive artifacts often depends on the kind of epistemic action (and its epistemic purpose) one performs with it. A shopping-list does not need to be very durable because after having bought the needed items, it has fulfilled its epistemic purpose. A computer, in contrast, *does* need to be durable because we need it for many kinds of epistemic actions for a long period of time. Wilson and Clark [16] introduce a trichotomy between one-offs, repeated, and permanent relationships with cognitive artifacts. Shopping-lists are one-offs. Abacuses or compasses, however, are frequently re-used because they are devices that are utilized many times for the purpose of calculating or navigating. But some cognitive artifacts enter into permanent and highly durable relationships with their users. Otto and his notebook, a carpenter and his slide ruler, and an astronomer and her telescope enter into long-lasting and interdependent relationships.

4.3 TRUST

In George Orwell's dystopian novel, *1984*, the ministry of truth continuously updates and changes information in entertainment, news media, and educational books with the purpose of rewriting history so that it fits the party's political doctrines. In addition to constant misinformation, people are persistently being monitored

by Big Brother and have therefore no privacy. So, if they are rational, people in this fictional world distrust (or ought to distrust) information that is provided and controlled by their government and should be very careful with what they write, publish, and distribute. Fortunately, in our non-fictional (Western) world things are better for at least two reasons. First, ideally we have freedom of press, freedom of speech, and freedom of information and thus control over the informational contents of our media and books. Second, we are not constantly being monitored by our government (although sometimes we are) and can thus write down, publish, and distribute whatever we desire. Freedom of information and informational privacy are two essential conditions for trust.

But there are other reasons for trust. Some information we trust because we have endorsed it somewhere in the past and wrote it down because of this. This is true for Otto's notebook, agenda entries, shopping-lists, and the information on post-it notes. Other information we trust because many people rely on it for their actions. This is true for timetables of trains, dictionaries, encyclopedias, and maps, which are used and shared by many humans. Because these symbol systems are shared with many others, and many people rely on them for their actions, there is often no reason to think that they are false or incorrect [18, contrast 19]. But there are exceptions: Wikipedia, for instance, is used and shared by many people, but is in some cases still not particularly trustworthy.

In two-way and reciprocal systems, we trust the information because we have endorsed it in the past and because we offloaded it ourselves, but we also trust it because we believe the information is private and has not been tampered with. Consider a brief example: in Australia there is a TV commercial for smart phones in which a parent goes shopping with a shopping-list composed on a smart phone. The application is connected in real-time to the desktop at home where his son deliberately changes the digital shopping-list to include items he desires. So with new digital cognitive artifacts with networking abilities such as smart phones and tablets, informational privacy and security [20, 21] become increasingly important for trust in information. Privacy and security issues are less likely to emerge when using analogue shopping-lists, which are identifiable by means of one's handwriting [22]. So the nature and properties of the artifact partly determine how relevant informational privacy and security are for establishing a trust relation with the artifact and the information it carries.

4.4 TRANSPARENCY

There are two types of transparency that are relevant for cognitive artifacts, namely, procedural and representational transparency. Embodied tools like bicycles, cars, hammers, and cricket bats transform the body schema. Body schemas are flexible as to incorporate tools into the sub-conscious representation of the body and its capabilities for action. Those tools, then, are not experienced as external objects with which one interacts, but one interacts with the environment through those tools [23]. When a tool is incorporated into the body schema of its user, it becomes transparent in use. We then no longer consciously need to think about how to interact with the tool, interaction goes smoothly and the tool withdraws from attention, i.e. it is transparent [24].

A similar thing happens with cognitive artifacts which I will call "procedural transparency". Procedural transparency [see also 25, 26] concerns the effortlessness and lack of conscious attention with which an agent deploys a cognitive artifact. Otto, for example, is so adapted and familiar to using his notebook that he will consult it automatically when he needs to do so. His perceptual-motor processes are proceduralized to such an extent that he does not consciously think about how to retrieve information from his notebook. So the retrieval process is not a two-step process in which Otto first believes that the address of MoMa is in his notebook and then looks up and interprets the information to form his second belief, namely, that MoMa is at 53rd street. Rather, it is a proceduralized and transparent process. In Clark's words: “the notebook has become transparent equipment for Otto, just as biological memory is for Inga" [26, p. 80]. Having a high level of procedural transparency needs substantial training and takes a considerable amount of time.

Representational transparency concerns the effortlessness with which an agent can interpret and understand external information. For example, in my neighborhood in Sydney there is a war memorial, *The El Alamein Fountain*, to remind us of the Australian soldiers that died in 1942 during the Second World War in El Alamein, Egypt. However, the memorial is a fountain and it is not immediately clear that it is meant to be a war memorial. Only after reading the commemorative plaque I understood what it is meant to represent. A fountain has very little, if any, representational isomorphism with war and casualties of war. So, for individuals who know the representational function of the fountain, it may evoke (strong) memories and emotions about the Second World War. Yet others who do not know its representational function, may perceive it as a mere aesthetic object and have no connection to what it represents. Thus, representational transparency is not an objective or intrinsic feature of cognitive artifacts, but partly depends on the cognitive profile and capacities of the interpreting agent. In contrast, *The Tomb of the Unknown Soldier* in Ottawa, Canada is functionally and representationally much more transparent, because it is comprised of a number of soldiers holding guns and has clearly and largely written "1914-1918" on a plaque placed under the soldiers. So for most people it is immediately clear that it is meant as a memorial for the First World War. Thus whether a memorial or other cognitive artifact fulfils its representational function partly depends on its representational transparency, which, in turn, partly depends on the degree of representational isomorphism.

4.5 INDIVIDUALIZATION & ENTRENCHMENT

Sterelny [18] has argued that certain cognitive artifacts are individualized and entrenched. For Sterelny, individualization is changing, adjusting, or fine-tuning the artifact such that its use is more effective and efficient. He argues that most of the books in his professional library are interchangeable, but some of them are massively individualized with underlining, highlighting, comments, and post-it notes. These adjustments essentially make sense to Sterelny and are less useful and valuable to others. Similarly, Otto's notebook is highly individualized and is useful only for Otto, although others may still be able to read the notebook, only Otto uses it to aid his memory and to structure his actions. My tablet computer is fairly individualized: it has applications that I have downloaded and installed to fit my specific needs such as the weather forecast and train timetables for Sydney, and specific websites, documents, and books. But although it is individualized, most applications are still easily usable by others. In contrast, no-smoking signs, road signs, and library books are not individualized (and thus interchangeable)

and accessible for most people. Individualization of cognitive artifacts often takes a certain period of time and highly individualized cognitive artifacts are in close equilibrium with the cognitive profile of their user.

Entrenchment of cognitive artifacts implies a close equilibrium between agent and artifact in which *both* have been transformed in order to ensure the best possible fit between agent and artifact. Sterelny acknowledges that his individualized books are not entrenched in the sense that his professional routines and habits have not been adjusted to those books in the same way as those books have been adjusted to Sterelny. So, he has individualized his books, but his books have not individualized him, or at least not sufficiently. But, according to Sterelny, there may still be cases of entrenchment concerning books. For a Locke scholar, Locke's oeuvre may have transformed the routines of the scholar in the same way as he or she has transformed Locke's oeuvre in the sense of highly individualizing his works by underlining, highlighting, comments, and so on. A more obvious and clear example of an entrenched cognitive artifact is Otto's notebook. The information in his notebook is only meant for Otto himself and is specifically geared to his needs and desires, so it is highly individualized, and his behavioral and cognitive routines are sculpted by his notebook, so it is entrenched as well.

4.6 BANDWIDTH

Like information flow in computer networks, information flow in cognition-artifact systems has a certain bandwidth, which is the amount of information that is exchanged per unit of time and depends on properties of both the agent and artifact. For example, a map of a city on which a particular route is outlined, potentially has a greater bandwidth than a linguistic description of the same route, because for most people it is easier and more effective to interpret a map, than to read a linguistic description of a given route. Similarly, a graph of the amount of carbon dioxide in the earth's atmosphere plotted against the time, a pie chart of the distribution of species in a given ecosystem, and an organization chart of the departments of a university make complex relationships between several items or variables clear and easily understandable. Graphs, pie charts, diagrams, and other illustrations transform an abstract relationship or problem space into a relatively easy to understand visual format. Explaining these relations in linguistic terms would in most cases be significantly more burdensome. In fact, this is often the point of using non-linguistic representations: to effortlessly and quickly convey information that would take much more time to explain in linguistic format. Common wisdom would say that a picture is worth more than a thousand words. Bandwidth also depends on the interpretative skills of the agent. Some agents can take more information onboard in a given amount of time than others.

4.7 SPEED OF INFORMATION FLOW

The speed with which information flows depends (again) both on the representational properties of the artifact and the cognitive profile of its user. Some people read quickly, while others do not. Some people interpret a map in one glance, while others have to study it before they know where to navigate. Humans have thus different interpretative skills, which partly determine how fast information is taken onboard and processed. The degree of representational transparency is also relevant here. Some information is easier to interpret than other. The higher the representational transparency, the easier the information is to interpret, and the higher the speed of information flow. So speed of information flow depends, on the one hand, on the cognitive and interpretation skills of the human agent and, on the other hand, on the informational and representational nature of the cognitive artifact. But contextual factors such as background noise may also influence speed of information flow, since one's concentration and thus also one's ability to interpret information is influenced by it.

Conversely, the speed with which one offloads information onto an artifact is also important. Again, this depends on properties of both the agent and artifact. Certain devices have input methods that are more efficient than others. A desktop computer has a keyboard that is geared towards quick data input, a tablet has a virtual keyboard that is much less efficient, and a smartphone has a virtual keyboard as well, but one that is much smaller and thus significantly less efficient. Some computing devices have auditory input methods which are potentially much quicker than conventional methods, because most people can speak quicker than they can type or write. But equally relevant are the interactive skills of the agent. Some people write or type considerably quicker than others, which often depends on training and education.

4.8 DISTRIBUTION OF COMPUTATION

The degree to which each element in a cognition-artifact system contributes to solving a problem depends on the distribution of computation. Compare, for example, making a graph by way of pen and paper with making a graph by way of a spreadsheet program. Let's assume that both graphs are based on the same dataset, so the cognitive output (i.e., the graph) is the same, but the distribution of computation is different. In the first scenario, most computation is performed by the human agent, whereas in the second scenario most computation is performed by the spreadsheet program. In the latter case, we delegate most of the information-processing to the artifact. The distribution of computation is relevant for the nature and coupling of the system, but of course only for artifacts that have themselves information-processing abilities. In case of analogue or static cognitive artifacts, all information-processing is done by the human component and the artifactual component then merely functions as a medium for storage with its own representational properties.

4.9 COGNITIVE & ARTIFACTUAL TRANSFORMATION

As we have seen, body schemas are flexible as to incorporate tools into the sub-conscious representation of the body and its capabilities for action. Tool-use thus transforms the body schema. Likewise, the use of cognitive artifacts and other external symbol systems transform the representational and cognitive capacities of the human brain. Helen de Cruz [27], Menary [28], Clark [9] and Michael Kirchhoff [29], amongst others, have argued that external symbol systems transform the brain's representational capacities. During ontogenetic development we interact with public representational systems such as mathematics and language. By so doing, we soak up and learn to think in those representational systems and the brain takes on the representational properties of those systems.

Language and mathematics are examples of external symbol systems with which we interact substantially for a long period of

time, both phylogenetically and ontogenetically. In ontogeny we call this period education. A considerable amount of research has been done on the transformation affect of those systems on our brain and cognition. Other cognitive artifacts and symbol systems such as road signs, maps, computers, and design programs have presumably also a transformation affect on our representational and cognitive capacities. For example, after navigating a city with a map for a certain period of time, the interaction with the map and the city has changed our internal spatial representation of parts of the city. At a certain point, we no longer need the map to navigate and we have to a certain degree internalized the map. Likewise, interacting with computers for many hours a day probably transforms our neuronal structures and cognitive capacities. Engineers, for example, spend many hours a day designing objects and structures with design programs. It is not unlikely that after a certain period of training and practice their brains take on the representational properties of the program. Such transformations seem to be a consequence of long-term interaction with cognitive artifacts over ontogenetic time.

It is, however, not only the human component of the system that transforms its representational properties and capacities. The artifactual component transforms its representational properties too: cognitive artifacts are often not static and fixed but active and dynamic. The representational properties of post-it notes, slide rulers, and textbooks, for instance, are fairly stable and fixed, but smart phones, tablets, laptops, and other computing devices are very dynamic in their representations. We can transform and adjust their representational properties to our own needs and desires, and it is frequently *because* we act on those artifacts and the information they carry that they have dynamic and changing representations.

5 DYNAMICS IN THE FRAMEWORK

All these dimensions are matters of degree and relational in the sense that they emerge out of a specific epistemic interaction between agent and artifact performed in a particular context and with a specific epistemic purpose in mind. Importantly, they are not meant as necessary conditions for cognitive extension and thus do not provide a clear set of conditions to demarcate between cases of embedded and extended cognition. On my view, particular cognition-artifact relations merely populate a certain region in this multidimensional space; the higher a specific cognition-artifact relation scores on these dimensions, the more tightly coupled the system is and the closer it is integrated with the human cognitive system. Let me now apply the above developed multidimensional framework to briefly analyze two distinct cognition-artifact relations.

First, when using a map during a citytrip, information flows from artifact to agent and so a one-way system is established. Say the agent consistently brings the map on each day of the citytrip, access to the information is thus highly reliable. The durability of the relationship is as long as the duration of the citytrip. So it is not a one-off or a permanent relationship, but a repeated one. The amount of trust in the correctness of the map is high, since the map was provided by an official travel agency. The procedural and representational transparency increases during the citytrip. The more often the map is deployed, the easier it becomes to use and interpret it. Let's assume that the agent does not make notes on the map, so it is not individualized. Both the bandwidth and speed of information flow depend on the representational properties of the map and the interpretative skills of the agent and are likely to increase over time. Maps that are simple and easy to interpret have potentially a greater bandwidth and speed of information flow. The distribution of computation is such that the agent does all the information-processing, because the map is a mere medium for information storage. And finally, depending on how often the map is deployed, it will (slightly) transform the representational properties of the agent. It is not unlikely that after a couple of days of navigating the city with the map, the agent partly transformed her internal representation of the city. But the map itself is static and does not transform after use. Thus, given how it scores on the above dimensions, this cognition-artifact relation populates a region somewhere in the middle of the space.

Second, Otto and his notebook constitute a two-way system: Otto offloads information (e.g., addresses, phone numbers, notes, ideas, et cetera) onto the notebook and then retrieves it for later use. Otto heavily depends on his notebook to successfully get around in the world, he therefore always carries it with him and thus the information in it is reliable available. A permanent relation is established in the sense that he consistently uses the notebook over a long period of time (Alzheimer's disease can take over a decade). He automatically trusts the information in the notebook, because he has endorsed it somewhere in the past and wrote it down because of this, but also because it is extremely unlikely that people will deliberately tamper with the notebook. For Otto, the notebook is highly transparent, both procedurally and representationally. Otto's perceptual-motor processes are proceduralized to such an extent that he does not consciously think about how to use retrieve information from the notebook. And because the information is written and structured by Otto himself, he does not need to think about what it means. For example, the sentence "MoMa is at 53rd street" needs little, if any, conscious deliberation. The notebook is further deeply entrenched, i.e. Otto has personalized his notebook, which, in turn, has sculpted his cognitive and behavioral routines and capacities. The bandwidth is fairly high, since the offloaded language is likely (though not necessarily) to be geared towards easy intake. The offloading speed is relatively fast, because Otto writes in his notebook and writing is a fairly quick method for offloading information. The distribution of computation is such that Otto performs all the information-processing, since notebooks are mere analogue mediums for information storage. And finally, the notebook may not have deeply transformed Otto's representational capacities, but language (i.e. the representational medium in his notebook) in general certainly has. Thus, given how it scores on the above dimensions, this cognition-artifact relation populates a higher region in the multidimensional space.

Further, existing relations can shift from one region to another. When a particular artifact is used for a longer period of time and it becomes gradually more individualized, transparent, and trustworthy, the relation between user and artifact becomes increasingly more integrated. As a result, the relation will shift to a higher region in the multidimensional space. Highly individualized and entrenched cognitive artifacts are likely to maintain a stable relation with its user and, consequently, populate a given region in the space for a long period of time, but most relations are constantly shifting from one region to another. This is so because most cognition-artifact relations are very dynamic in nature, constantly changing their functional and

representational properties, and renegotiating existing informational equilibriums.

For analytical purposes, I have discussed each dimension separately, but some of them overlap and interact. I shall now very briefly look at a number of these interactions. Reliable access and durability often result in individualization. The more often a certain cognitive artifact is used, the more likely it is that it will be individualized and perhaps in some cases even entrenched. But this need not be the case. There are often-used cognitive artifacts that are not individualized or entrenched such as clocks and speed dials. Individualization and entrenchment frequently result in cognitive transformation. Again, the more often we use a certain cognitive artifact, the more likely it is that the human brain takes on the representational properties of the artifact. This happens with language and mathematics, but also with maps, design programs, and perhaps with graphs, pie charts, diagrams and other illustrations as well. Individualization frequently causes both trust and procedural and representational transparency. Individualized cognitive artifacts, including agendas and notebooks, are designed by the user of the artifact and thus almost automatically trusted and transparent in use, as well as transparent in interpretation. We do not need to think about how to use such artifacts, and the information they carry is trusted because we wrote it down ourselves. And finally, representational transparency often results in a higher speed of information flow. The idea being that the easier information is to interpret and understand, the faster we can take it onboard and process it. There are more interactions between the dimensions, but these are the most obvious ones.

6 CONCLUSION

This paper first briefly discussed the concept of a cognitive artifact and then distinguished between three levels of functional and informational integration between human agents and cognitive artifacts, including monocausal and bicausal relations as well as continuous reciprocal causation. After that, a multidimensional framework for exploring cognition-artifact coupling was sketched. Collectively, the dimensions constitute a multidimensional space in which cognition-artifact relations can be located. The framework provides a toolbox for detailed studies of specific conceptual or empirical cases of the use of cognitive artifacts. The higher a cognition-artifact relationship scores on these dimensions, the higher a region in this space it will populate, in which case there is higher degree of integration.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisors John Sutton and Richard Menary as well as two anonymous reviewers for valuable and perceptive comments on an earlier version of this paper.

REFERENCES

[1] R. Heersmink, 'Defending Extension Theory: A Response to Kiran and Verbeek', *Philosophy and Technology*, doi: 10.1007/s13347-011-0035-6, (2011).

[2] M. Rowlands, 'Extended Cognition and the Mark of the Cognitive', *Philosophical Psychology,* 22, 1-19, (2009).

[3] A. Clark & D. Chalmers, 'The Extended Mind', *Analysis,* 58, 7-19, (1998).

[4] M. Wheeler, In Defense of Extended Functionalism, In: *The Extended Mind,* R. Menary (Ed.), MIT Press, Cambridge, (2010).

[5] M. Wheeler, 'In search of clarity about parity', *Philosophical Studies,* 152: 417-425, (2011).

[6] J. Sutton, Exograms and Interdisciplinarity: History, the Extended Mind, and the Civilizing Process. In: *The Extended Mind,* pp. 189-225, R. Menary (Ed.), MIT Press, Cambridge, (2010).

[7] J. Sutton, C.B. Harris, P. Keil, & A.J. Barnier, 'The Psychology of Memory, Extended Cognition, and Socially Distributed Remembering', *Phenomenology and the Cognitive Sciences,* 9, 521-560, (2010).

[8] R. Menary, Cognitive Integration: Mind and Cognition Unbounded, Palgrave McMillan, (2007).

[9] A. Clark, Being There: Putting Brain, Body and World Together Again, MIT Press, Cambridge, (1997).

[10] J. Kiverstein & M. Farina, 'Embraining Culture: Leaky Minds and Spongy Brains', *Teorema,* 32, 35-53, (2011).

[11] D.A. Norman, Cognitive Artifacts, In: *Designing Interaction: Psychology at the Human-Computer Interface,* pp. 17-38, J.M. Carroll (Ed.), Cambridge University Press, (1991).

[12] K. Beach, The Role of External Mnemonic Symbols in Acquiring an Occupation, In: *Practical Aspects of Memory: Current Research and Issues Vol. 1,* pp. 342–346, M.M. Gruneberg & R.N. Sykes (Eds.), Wiley, New York, (1988).

[13] E. Hutchins, Cognitive Artifacts, In: *The MIT Encyclopedia of the Cognitive Sciences,* pp. 126-128, R.A. Wilson & F.C. Keil (Eds.), MIT Press, (1999).

[14] R. Menary, 'Writing as Thinking', *Language Sciences,* 29, 621-632, (2007).

[15] J.L. McClelland, D.E. Rumelhart & G.E. Hinton, The Appeal of Parallel Distributed Processing, In: *Parallel Distributed Processing, Volume 2,* pp. 3-44, J.L. McClelland & D.E. Rumelhart (Eds.), MIT Press, (1986).

[16] J. Sutton, 'Distributed Cognition: Domains and Dimensions', *Pragmatics and Cognition,* 14, 235-247, (2006).

[17] R. Wilson & A. Clark, How to Situate Cognition: Letting Nature Take its Course, In: *The Cambridge Handbook of Situated Cognition,* pp. 55-77, M. Aydede and P. Robbins (Eds.), Cambridge University Press, (2009).

[18] K. Sterelny, 'Minds: Extended or Scaffolded?', *Phenomenology of the Cognitive Sciences,* 9, 465-481, (2010).

[19] K. Sterelny, Externalism, Epistemic Artefacts and the Extended Mind, In: *The Externalist Challenge. New Studies on Cognition and Intentionality,* pp. 239-254, R. Schantz (Ed.), De Gruyter, (2004).

[20] L. Floridi, 'The Ontological Interpretation of Informational Privacy', *Ethics and Information Technology,* 7, 185–200, (2005).

[21] L. Floridi, 'Four Challenges for a Theory of Informational Privacy', *Ethics and Information Technology,* 8, 109-119, (2006).

[22] M. Parsell, 'The Cognitive Cost of Extending an Evolutionary Mind into the Environment', *Cognitive Processing,* 7, 3-10, (2007).

[23] R. Heersmink, 'Embodied Tools, Cognitive Tools, and Brain-Computer Interfaces', *Neuroethics,* doi: 10.1007/s12152-011-9136-2, (2011).

[24] M. Merleau-Ponty, *Phenomenology of Perception*, Routledge, (2003).

[25] J. Sutton, The Feel of the World: Exograms, Habits, and the Confusion of Types of Memory. In: *Philosophers on Memento,* pp. 65-86, A. Kania (Ed.), Routledge, (2009).

[26] A. Clark, Supersizing the Mind: Embodiment, Action, and Cognitive Extension, Oxford University Press, New York, (2008).

[27] H. de Cruz, 'An Extended Mind Perspective on Natural Number Representation', *Philosophical Psychology,* 21, 475-490, (2008).

[28] R. Menary, 'Dimensions of Mind', *Phenomenology and the Cognitive Sciences,* 9, 561-578, (2010).

[29] M.D. Kirchhoff, 'Extended Cognition and Fixed Properties: Steps to a Third-Wave Version of Extended Cognition', *Phenomenology and the Cognitive Sciences,* doi: 10.1007/s11097-011-9237-8, (2011).

Turing and the Real Girl

Dr Stephen Rainey[1] and Dr Yasemin J. Erden[2]

Abstract: Alan Turing's oft-cited remark about the possibility of machine thought, and its relevance for machine intelligence or even agency, continues to provoke interdisciplinary debate about the nature of such terms. This is in particular regard to the likelihood that the Turing Test could actually solve questions about machine intelligence. In this paper we centre our discussion on these topics by focussing on the complexity of identity or personhood in terms of agency. We do this by exploring concepts such as shared communication, recognition, and wider forms of validity. The crux of the paper is this: Alan Turing asked in his seminal paper whether machines could think. To this we add: Would we be willing to recognise it as thought even if it did?

1 INTRODUCTION

Alan Turing's oft-cited remark about the possibility of machine thought [1], and its relevance for machine intelligence or even agency, continues to provoke interdisciplinary debate about the nature of such terms. This particularly concerns the likelihood that the Turing Test could actually solve questions about machine intelligence. In this paper we explore this question by focussing on the complexity of identity or personhood. In order to do this, we approach the topic in two ways. First, we discuss the idea of agency, and in particular agency with regards to action. More specifically, we consider what is understood by the term 'actor'. Second, we engage with the idea of recognition as an essential, yet underdeveloped perspective from which to view or engage with the idea, and more importantly, how we come to ascribe intelligence to other actors, and in whether and how we might do so for machines. Before we can follow this train however, we need to look more closely at Turing's ideas.

First we need to be clear that Turing's initial question, about whether machines can think, is extremely problematic, and for a number of reasons. In the first instance this is because he does not actually focus on this question in his paper and instead conflates it to one concerned with imitation [1]. This pivotal decision has important ramifications. In fact, as we argue throughout, what we are willing to accept as *thought* and what we are willing to accept as *imitation* are not synonymous. One could accept that a machine could succeed in the imitation game, yet not accept that the machine is thinking, without engendering a contradiction. Accepting the first premise would be equal to accepting that the machine can 'act'. The imitation game does not prove that there is agency, or that there is an agent, in the sense of something, or someone, that we may be willing to recognise as an agent, with agency. This is as opposed to an actor who may be able, even successfully, to imitate agency.

[1] FUNDP, Belgium
✉ stephen.rainey@fundp.ac.be
[2] St Mary's University College, London, UK
✉ erdenyj@smuc.ac.uk

Turing's initial question 'can machines think?' immediately leads us to ask 'what do we mean by the words *to think*?' As Turing identifies, this question is almost impossibly difficult to answer, where the answer is sought in the normal usage of the words [1]. In a similar vein Wittgenstein describes that language is only meaningful within the context of language-games: 'The speaking of language is part of an activity, or of a life-form—this is what "language game" brings into prominence. Language is communal' [2, §23]. It is for these reasons that Turing decides to take a different tact and instead focus on the 'imitation game':

> We now ask the question, 'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, 'Can machines think?' [1, p. 433]

The problem is that, as already observed, these questions are markedly different, and would in fact solve different problems. To pass the Turing test requires that imitation be successful *as just imitation*, such that we would be willing to accept that in this instance there was the appearance of agency, rather like we may also accept the genuine agency of actors on a stage. Accepting that there is thought on the other hand requires more than just appearance. This requires participation in a form of life, as well as recognition within that form of life. Wittgenstein identified this as a problem when he stated: 'it isn't true that we are never certain about the mental processes in someone else. In countless cases we are' [3, p. 94]. To which he adds:

> And now the question remains whether we would give up our language-game which rests on 'imponderable evidence' and frequently leads to uncertainty, if it were possible to exchange it for a more exact one which by and large would have similar consequences. For instance, we could work with a mechanical 'lie detector' and redefine a lie as that which causes a deflection on the lie detector.
>
> So the question is: Would we change our way of living if this or that were provided for us?—And how could I answer that? [3, p.95]

The empirical difficulty of answering a question of that sort is detailed elsewhere when he imagines the impossibility of communication with a lion, but there is a further difficulty, which involves agency. Namely: would we want to even give such ground? In this vein, would we be willing to accept truth as mediated by a lie detector, and even if in some circumstances we may (think of a police officer trying to gain evidence from a suspect), does this mean we would accept this manner of discovering truth above our standard language-games that rest on 'imponderable evidence'. What Wittgenstein highlights here is that agency includes choice, as situated within a form of life. For

instance, would precision be what we accept as most important in all situations? To answer this, there is another quotation we might consider:

> Just think of the words exchanged by lovers! They're 'loaded' with feeling. And surely you can't just agree to substitute for them any other progressions of sound you please. Isn't this because they are *gestures*? And a gesture doesn't have to be something innate; it is instilled, and yet *assimilated*.--But isn't that a myth?!--No. For the signs of assimilation are that I want to use *this* word, that I prefer to use none at all to using one that is forced on me, and similar reactions. [3, p.17]

The above would seem to lay further doubt on Turing's attempt to replace one question with another in the way that he does, since it assumes that the belief by agents that other agents have thought could be equal to the imitation of thoughtful action by a machine. Imitation need not be accepted as agency, and that is why we need not accept apparent thinking as indeed thinking. To defend our claim further, we now offer a more detailed account of what is meant by agency.

2 AGENCY, ACTION, ACTOR

Personhood and agency are terms with pedigree both within philosophy and within computing. As complex terms they imply attributes (to a greater or lesser degree) such as autonomy, intelligence and intentionality, amongst others. Such attributes are typically valued in the development of intelligent computing systems, not least since the possibility of autonomy is often seen as essential for particularly sophisticated Multi Agent Systems [4]. Philosophical discussion about these terms is expansive but for our purposes here, a short review will suffice. For Kant, autonomy is conforming with universal laws revealed by reflection. For Hume, autonomy involves thinking about sentiment -- what sort of a character one wants to be. For others such as Searle or Brandom, having reasons from various sources private and public, permit autonomous action. For yet others such as Wittgenstein, acting autonomously means involving oneself in a general way of being, as engagement within a form of life [2, 5, 6, 7, 8].

While these accounts of autonomy are perhaps familiar and somewhat intuitive, philosophical discussion regarding the ascription of action in terms of 'acting' is rather less so. In ordinary terms it is quite natural to think of 'acting' as having some association with pretence or deception, or with the intention to deceive. In this way the idea of an 'actor' may equally be associated with the identity of a *pretender*. In common parlance, we might think about, for example, dramatic portrayals on the stage or screen. If we set our sights on creating a machine capable of acting in this sense, it would seem reasonable to accept that we would only need to create something that is superficially similar to its target: an automaton going through the motions, as it were. As we show above objections to the Turing test can cluster around this caveat, that the passing of the test requires a behaviouristic reduction that nevertheless remains both superficial and/or unsatisfying.

In a philosophical sense, however, 'actor' carries more import than this: an actor is a *locus of actions*. Action here contrasts with something like a reflex, since reason is implicated in action in a way that it is not in reflex. Raising my knee following a tap with a hammer in the doctor's surgery is not an action in the same way that my making the same movement to take a shot at goal would be. While the former is purely a physiological response, the latter is considered an action because I do something more than automatic response. As such, action engenders concepts of intentionality and choice: it involves my *being an actor*. Being an actor can mean being sensitive to reasons, since in the absence of reasons, nothing can be considered as intentional or deliberate. In short, nothing can be considered as an action where there is not the intention or will to act. In a world without reasons, we have a world without *action*.

As already noted this account of *action* as opposed to reflex or automatic behaviour puts into question the requirements by which a machine or computer might be thought capable of passing the Turing test. This means that even if a machine were able to do so, we might consider this notion of acting or action only in the sense of imitation or pretending, namely that the machine goes through the motions of exhibiting intelligent behaviour, or 'having' intelligent thoughts. In order to manifest Turing-Test-passing behaviour, by which we mean purposive behaviour akin to action, we might suppose that one would need only to have *cause* to do either this or that. On this it would be fair to acknowledge that the creation of a store of causes, such that superficially quite complex behaviour could be manifested, would not need to be more complex than a program that generates rules for the processing of input and output protocols. Nevertheless it would be difficult to conclude that this would be sufficient as an account for the kind of deliberative action we have just outlined.

It is important to note at this stage that the preceding discussion, regarding intentional or deliberative action, also involves the notion of choice: in a reflex action, I have little or no choice. Yet when a striker shoots to score it is, at least in substantial part, *because he wants to.* The point is that there are choices to be made rather than protocols to be followed. The rigidity of a protocol might well limit decision making to a single subset of reasons, but reasons are not all vanilla. Basic insensitivity to a selection of intention-driven, subtle yet complex reasons limits potential for action in the intentional sense that we identify above. So another aspect of this idea of action involves the ability to select from and apply a broad range of reasons, but in ways that are sometimes, perhaps often, unquantifiable, especially before the event (sometimes even after). What is crucial for our primary purpose in this paper is that in the *recognition* of these broad bases the actor is thereby *recognisable* as a reason-user. The process is necessarily intersubjective. Indeed many reasons occur only in virtue of intersubjective relations, and many intentional actions can only succeed in virtue of those relations. Power relations are a good example here: giving commands requires the commander and commanded to be suitably related interpersonally, with shared understanding of reasons (whether these reasons are shared or not is another matter), in order to expect compliance, as commander, or to follow orders, as the commanded.

Yet intelligent action does not engender the perception of either agency or identity, even in recognisably human forms. These elements are little guarantee of acceptance as a person or personhood. There are in fact political and social implications of the concept of, or indeed need for, *recognition*. This was often

one of the key defences offered for slavery, namely the non-recognition of the personhood of humans identified as "other", despite clear evidence to the contrary. What this highlights, and in line with that noted by Hegel in his chapter on *Master and Bondsman*, is that the recognition of personhood is by no means always a simple matter of the evidence of intelligence. In concentrating upon the technical possibilities of machine intelligence and the philosophical responses to these we overlook a broader philosophical point about the nature of identity as hinging upon this idea of recognition. As Hegel famously remarks 'self-consciousness exists in and for itself when, and by the fact that, it so exists for another; that is, it exists only in being acknowledged or "recognised"' [9, §178]. To this, Redding adds: 'we are the sorts of beings we are with our characteristic "self-consciousness" only on account of the fact that we exist "for" each other or, more specifically, are *recognized* or *acknowledged* (*anerkannt*) *by* each other, an idea we might refer to as the "acknowledgment condition" for self-consciousness' [10].

This latter concept provides illuminating insight into personal identity and suggests a relatively unexplored dimension of the debate on machine intelligence—the role of relations in the world among interlocutors as a condition of acceptability (although it is touched on in [4], it is not the primary concern). The question then becomes what they could mean for recognition of personhood in non-singularly-biological entities. The point is that, natural or synthetic, concepts like agency, intentionality and intelligence are difficult to pin down. It is clear that there are important distinctions to be drawn here among these concepts with respect to the issue of personal identity. It is also apparent that such distinctions will affect what we expect the Turing Test to resolve. As such, we would do well to disentangle these issues while also exploring their points of contact.

We conclude this section with an important clarification regarding our purpose. While we maintain, contra most functionalist positions, that physical differences between machines and humans are not insignificant, the purpose of this paper is not to enter into that debate. There has already been significant argument in these areas, and it is clear that further discussion will be required on both sides of the discussion, yet this is not our primary interest here. Instead, the problematic we seek to explore is that associated with the recognition of an agent as an agent. The contention of our paper is that such recognition is required should an agent be considered as an agent in a full sense of the word: capable of intentional action and participation in a form of life, for instance. It is evident from history that evidence of thought does not secure this recognition and so even in clear cases of imitation of thought (whether genuine or not) there is not yet the basis to ascribe agency. Agency is not reliably conferred on any particular behavioural basis at some particular moment, but rather arises across a range of intersubjective experiences.

As already briefly noted above, reasons to deny recognition come from various sources: history, culture, judgements about character, bigotry, economic status, and the list goes on. An apparently thinking machine could on a similar rationale be excluded in virtue of its silicone nature. An apparently thinking programme could be excluded in virtue of its having been assembled in multiple instances. The point is that factors extrinsic to the person, machine or programme feature in its determination as an agent. Agency is not simply a facultative matter, but draws upon the will of others to be constituted. Again, this is not our primary concern.

In short, our argument here is that intelligence, even when evidenced in deep, involved and complex ways, nonetheless need not be considered co-extensive with the notion of personhood, in the sense that intelligence is not a necessary nor a sufficient precursor to personhood. As such, even if we were to develop an incontrovertibly intelligent machine, the question would remain as to the *recognition* of the machine (in the first instance by us, but also in terms of mutual reciprocity) as something person-like, let alone as a person. Recognition and acceptance are, we suggest, shown to be key elements in both personhood and identity. The potential for machines to be recognised as equivalent to persons (in the way that some marine biologists might wish for dolphins, for example, to also be recognised as persons, or as having personhood) remains to be seen. Our point is that intelligence will not offer a shortcut to these criteria. The fact that Turing begins his paper with the questions of whether machines can think is, we suggest, part of the problem, since the move from this to mimicry and the conflation of thinking with intelligence, has led to an over-simplification of personhood and agency.

One key way in which we identify personhood is through communication, and primarily language. It is no coincidence that Turing's test requires that there is zero embodied communication between interlocutors, since that, it is thought, would give the game away. Recognition between agents centres on the interpersonal relationship between them, which in turn centres on the possibility for communication with a shared language. It is to this idea that we now turn.

3 LANGUAGE, COMMUNICATION AND VALIDITY

A constant in communicative understanding is the presence of interlocutors *qua* interpreters [11]. It is not simply the sharing of theories (having a common, pre-learned language) or common adherence to conventional practices that allows for us to understand one another. More than this, it is that we can interpret one another and be interpreted as acting on or according to reasons. Such mutual interpretation will succeed or fail on the basis of the ingenuity of the interpreters, the relative familiarity of the circumstances, and various bits of collateral knowledge: For example, despite the promptings of even Ziggy and Al, it always takes Sam Beckett of the cult sci-fi programme *Quantum Leap*, half an episode of engaging with the rest of the protagonists before he can figure out what is going on (and hopefully secure the next leap as the final leap home). Donald Davidson might conclude from a scenario such as this:

> …that there is no such thing as a language, not if a language is anything like what many linguists and philosophers have supposed. There is therefore no such thing to be learned, mastered or born with. We must give up the idea of a clearly defined shared structure which language-users acquire and then apply to cases. [12, p. 174]

One consequence of this account is that it seems more difficult to accept linguistic expressions made by machines as significant,

by which we mean as *meaningful in and of itself by virtue of its intentionality*.[3] Instead we are drawn back to the familiar conclusion that such expressions do in fact amount to little more than empty syntax manipulation that contains no semantic intention on the part of the machine, and does in fact gain significance only by way of our involvement with and thereby also our response to such expression. It is on this claim that Searle bases his now infamous Chinese Room example, about which there is substantial discussion in Preston and Bishop [14]. Suffice to say that for our purposes in this piece we wish to draw upon the ideas of interpretation and the resources used in interpreting. We will do this in order to pursue our broad agenda of highlighting issues beyond the merely technical regarding intelligence, action, agency and so on.

In devising his test Turing drew upon the intuitions of thinkers as diverse as Liebniz, Gödel and Frege to attempt to develop a calculation system that could account for human reasoning [15]. Frege's logical developments, such that permitted first and second order propositional calculus, cf. [16], represent a crucial point of departure in this scheme, and so acts as a fundamental moment in the development of artificial reasoning—the contention here is that this crucial moment also provides clues to some intrinsic challenges to the idea. These challenges arise because human beings reason in ways that draw upon a range of sources, not all of which can be easily coded into a propositional form, nor can they be easily identified or established. These can include power relations, senses of validity not limited to logical validity, and often require the use of a notion of speech action. The word 'act' is important here, particularly because of the impetus given to action above. Action, as noted already, points to a 'reasons based' behaviour, and not behaviour that is merely derived from premises according to narrow conceptions of valid inference. The scope of reasons for familiarly *reasonable beings* is much wider than the model shown in formal logic. This question of the scope of validity is not one that is easily resolved.

'Validity,' in a communicative sense, is an issue tackled by Jürgen Habermas, who develops ideas about the philosophy of language of the 1950s. The validity basis of speech hinges upon the fundamental thought that in the very act of uttering, a speaker is claiming to be:

> - giving [the hearer] *something* to understand
> - making *himself* thereby understandable; and
> - coming to an understanding *with another person*.

[17, p. 119]

These are three 'world relations' taken to be implicit in speech action; fully successful speech acts must satisfy conditions of truth, sincerity and accountability (i.e. legitimacy according to some specifiable criteria) [18, p.75]. An assertion *that p* raises the claim that p is true; that my utterance thereof is sincere, is an accurate reflection of what I believe; that my uttering of p is appropriate in the circumstances. Similarly, if I command you to bring me a drink my command raises the claims that there is drink to be had close at hand; that I would actually like a drink; that I am in a position to issue commands (I am your superior in an appropriate respect, say).

Truth claims are not privileged as the only or the most important claims that can be raised in communication. A range of 'validity spheres' that also provide discursive space for claims to moral rightness, ethical goodness, authenticity or personal sincerity, and aesthetic value are discussed [18, p. 23]. Habermas supposes that claims in each of these spheres can be raised and redeemed in communicative encounters, which amounts to raising and redeeming claims by means of argument. This being the case, validity in spheres beyond that of truth can be thought of as involving a notion of *correctness* appropriate to their own standards as truth is appropriate to claims of factual accuracy.

In this context, the phrase "validity claim," as a translation of the German term *Geltungsanspruch*, does not have the narrow logical sense (truth-preserving argument forms), but rather connotes a richer social idea—that a claim (statement) merits the addressee's acceptance because it is justified or true in some sense, which can vary according to the sphere of validity and dialogical context [19]. Validity claims are thus symbolic or explicitly made defensible propositions, sensitive to context; 'A *validity claim* is equivalent to the assertion that the *conditions for the validity* of an utterance are fulfilled' [18, p. 38].

This is clearly not limited to the validity in which logicians are primarily interested. Rather, this must be taken to include in its scope myriad subtle nuances and relations in the world. Given we are in communication and not in some way merely noting one another's utterances, or in a therapy session or some other special type of interaction, we have to expect to be able to assume boundaries that themselves engender questions clustering around the themes of truth, sincerity and accountability. Given that these are neither quantifiable nor predictable, it is difficult to see how they might be accounted for within a system that does not, indeed cannot act in a way such that we would be willing to see its action as anything more than *acting* in the standard sense of that word. Nevertheless, it is also true to say that the difficulty with deciding the parameters for recognition are also irresolvable. It seems therefore that we arrive at a precipice whereby we must judge how far the act of recognition itself might resolve this tangled web of problems.

Habermas' three world relations are, at least, dimensions of appreciation or critique that can attend any locution: truth, sincerity and accountability. This means that, whatever the utterance, we can question its veracity, the will behind its use and the right to say it. In cases of truth, matters are quite straightforward. Verification, falsification and problematisation can be sought. In matters of sincerity, we draw upon more complicated notions including trust, familiarity with the speaker, the possibility of irony, bad faith and so on. In the last instance, things are even more tricky: we seek to establish whether the speaker can offer us some warrant that they are in a position to make the utterance they do.

For example, as noted above, if they are commanding us to do something, that they are suitably superior to us or otherwise have some manner of legitimate power over us. In these last two dimensions of possible critique it is abundantly clear that nothing internal to the utterance will give us more than just a clue as to how to go on. Ambiguities aside, the semantic content of the utterance might be clear enough, but the significance of the utterance as it is uttered by the speaker in the context it appears draws upon the recognition of these intentional and power relations that exist and persist among individuals. The

[3] There is much scope for discussing the various roles of intention, interpretation and reason in accounting for linguistic understanding that lies beyond the scope of this paper. For an overview, see the opening sections of [13].

significance is settled, to the extent that it can be, in concert with others. This can, in fact, be read as an analysis of the concept of 'being taken seriously,' which must be a *sine qua non* for personal agency. Recalling Hegel, from whom much talk of recognition stems, we can see this notion of interdependence in the description of how master and bondsman are related. We might assume, for instance, that Master 'exists only for himself' since 'that is his essential nature', and yet, Hegel tells us:

> he is the negative power without qualification, a power to which the thing is naught. And he is thus the absolutely essential act in this situation, while the bondsman is not so, he is an unessential activity. But for recognition proper there is needed the moment that what the master does to the other he should also do to himself, and what the bondsman does to himself, he should do to the other also. On that account a form of recognition has arisen that is one-sided and unequal. [9, §191]

It is for this fact that the recognition necessary for the Master's true independence is also necessarily lacking. Despite the initial *appearance* of independence, 'he really finds that something has come about quite different from an independent consciousness. It is not an independent, but rather a dependent consciousness that he has achieved.' [9, §192].

4 CONCLUSION

When we suppose that we develop technology with levels of intelligence, action or autonomy either equivalent to, or surpassing our own, we need to be clear about what it is that we expect such developments to achieve. Technical debates about the possibility of machine intelligence, about simulation versus instantiation of consciousness, may not resist solution, but the possibility for personhood is far more elusive. Discussion of this more elusive concept is fascinating, complex and holds importance for technical researchers and philosophers alike. But the concern raised here, is that of the criteria by which human beings might accept artificial agents as agents is at once complex and unknown, and relies on more than just the identification of intelligence or intelligent behaviour. Objective measures are not the only factors in determining whether a form of life is accepted as such.

The central and often unexplored issue that we have discussed here concerns the responsibility that must be undertaken by any putative 'us' should we wish not to foreclose on the possibility of expanding the scope of 'agent' to include machine intelligence. In this discussion, we considered the Turing test in terms broader than those normally used. We approached it as if it were akin to a *citizenship test*. Our point was that no matter the success or failure in the event, it is not a guarantee that human beings would accept it as any more than pretend action. The task is not just to design a system that can 'act human', since being human can not be boiled down to a series of behaviours that is nevertheless no more than a technical system that pretends to be human. These sorts of behaviour are relevant, certainly, just as the fact that we sometimes react reflexively. But beyond that, there is an element of judgement required among those who would be the peers of, say, a silicon being, and would in part involve what they are willing to recognise as action, what they will accept into their form of life, and what would feature in their spheres of validity. Just as the community of which Lars was a member, took time to decide whether, and how, his silicon girlfriend would be accepted as one of them (*Lars and the Real Girl*). There is indeed a technical problem to surmount, this is true, but that does not provide necessary or sufficient conditions that can prompt these much subtler judgements.

ACKNOWLEDGEMENTS

We would like to thank the referees for their comments, which helped improve this paper.

REFERENCES

[1] Turing, A. M., Computing Machinery and Intelligence, *Mind*, New Series, Vol. 59, No. 236 (Oct., 1950), pp. 433-460
[2] Wittgenstein, L. *Philosophical Investigations*, 2003, Oxford: Blackwell.
[3] Wittgenstein, L. *Last Writing on the Philosophy of Psychology: The Inner and the Outer, Volume 2*, Oxford: Blackwell.
[4] Magill, K., and Erden, Y. J., Autonomy and desire in machines and cognitive agent systems, *Cognitive Computation*, 2012, DOI DOI 10.1007/s12559-012-9140-9
[5] Kant, I., *Groundwork for the Metaphysic of Morals*, 1785 (any edition)
[6] Hume, D., *A Treatise on Human Nature*, 1739 (any edition)
[7] Searle, J., *Rationality in Action*, 2001, MIT Press
[8] Brandom, R., *Making it Explicit*, 1994, Harvard University Press
[9] Hegel, G. W. F. *The Phenomenology of Spirit*, §178, 1807 (Any edition)
[10] Redding, P. (2008). The Independence and Dependence of Self-Consciousness: The Dialectic of Lord and Bondsman in Hegel's Phenomenology of Spirit from *The Cambridge Companion to Hegel and Nineteenth-Century Philosophy*, Cambridge: CUP. pp. 94 – 110.
[11] Talmage, C., Davidson and Humpty Dumpty, Nous, Volume 30, pp. 537-544
[12] Davidson, D., A Nice Derangement of Epitaphs, in *Philosophical Grounds of Rationality*, Grandy & Warner (Eds.), OUP, 1986.
[13] Rainey, S., Austin, Grice and Strawson: Their shadow from Pittsburgh to Frankfurt," *Essays in Philosophy*: Vol. 8: Iss. 1, 2007 Article 17.
Available at: http://commons.pacificu.edu/eip/vol8/iss1/17
[14] Preston, J. and Bishop, M., *Views Into the Chinese Room: New Essays on Searle and Artificial Intelligence*, 2002, Oxford: OUP
[15] Davis, M., *The Universal Computer: The Road from Leibniz to Turing,* 2000, New York: Norton
[16] Frege, G., On Sense and Reference, translated by Max Black, Philosophical Review 57, 1948, pp. 207-30, reprinted in Peter Geach and Max Black (eds.), *Translations from the Philosophical writings of Gottlob Frege,* 1960, Oxford: Blackwell, second edition
[17] Habermas, J., What Is Universal Pragmatics? in *The Habermas Reader*, (Outhwaite, W., Ed), 1996, Cambridge: Polity Press 1996
[18] Habermas, J,. *The Theory of Communicative Action*, Vol. I, 1981, Boston: Beacon Press
[19] Bohman, J. and Rehg, W. "Jürgen Habermas", *The Stanford Encyclopedia of Philosophy (Spring 2008 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2008/entries/habermas/>. [accessed 10 May 2012]

Weak vs. Strong Computational Creativity

Mohammad Majid al-Rifaie[1] and **Mark Bishop**[2]

Abstract. In the spirit of Searle's definition of weak and strong artificial intelligence, this paper presents a discussion on weak computational creativity in swarm intelligence systems. It addresses the concepts of *freedom* and *constraint* and their impact on the creativity of the underlying systems. An analogy is drawn on mapping these two 'prerequisites' of creativity onto the two well-known phases of exploration and exploitation in swarm intelligence algorithms, followed by the visualisation of the behaviour of the swarms whose performance are evaluated in the context of arguments presented in the paper.

1 INTRODUCTION

In recent years, studies of the behaviour of social insects (e.g. ants and bees) and social animals (e.g. birds and fish) have proposed several new metaheuristics for use in collective intelligence resulted from social interaction.

Among the many works in the fields are research on swarm painting (e.g. [24, 7, 34, 35]), ant colony paintings [19, 23, 31]) and other multi-agent systems (e.g. RenderBots [29] and the particle-based non-evolutionary approach of Loose and Sketchy Animation [15]).

In most of the swarm-based work mentioned above (e.g. [24, 7, 34, 35, 19]), the painting process does not re-work an initial drawing, but rather focuses on presenting "random artistic patterns", somewhere between order and chaos [35]. Other classes of research (e.g. by Schlechtweg et al. [29] and Curtis [15]) are based on reworking an initial drawing. There is a significant number of related papers in the area of non-photorealistic rendering; particularly, many papers approach drawing and painting using the optimisation framework. Furthermore, particles have been used for stippling and other aesthetic styles in numerous papers. Turk and Bank's work [33] is an early example of optimising particle positions to control a stroke-based rendering. Hertzmann [21] optimised a global function over all strokes using a relaxation approach. In one of his works, Collomosse [14] used a global genetic algorithm to define a rendering algorithm. More recently, Zhao et al. [38] deployed an optimisation-based approach to study the stroke placement problem in painterly rendering, and presented a solution named stroke processes, which enables intuitive and interactive customisation of painting styles.

This work is an extension of ideas first presented at the Computing and Philosophy symposium at AISB 2011 [1] and subsequently published in the Cognitive Computation journal [6]. In the work discussed herein the impact of freedom and constraint on the concept of 'creativity' is discussed, followed by a discussion on the creativity of swarm intelligence systems. This paper also addresses the issue of weak verses strong computational creativity.

[1] Department of Computing, Goldsmiths, University of London, UK, email: m.majid@gold.ac.uk
[2] Department of Computing, Goldsmiths, University of London, UK, email: m.bishop@gold.ac.uk

2 ON CREATIVITY, FREEDOM AND ART

For many years there has been discussions on the relationship between art, creativity and freedom; a debate elegantly encapsulated in the famous German prose by Ludwig Hevesi at the entrance of the Secession Building in Vienna:

> *"Der Zeit ihre Kunst*
> *Der Kunst ihre Freiheit"*

That is: "To Time its Art; To Art its Freedom".

Which, centuries after, resonates an earlier observation from Aristotle (384-322 BCE) [17] emphasising the importance of freedom (here, having "a tincture of madness") in presenting a creative act.

> *"There was never a genius without a tincture of madness."*

On the other hand Margaret Boden, in [9], more recently argues that creativity has an ambiguous relationship with freedom:

> *"A style is a (culturally favoured) space of structural possibilities: not a painting, but a way of painting. Or a way of sculpting, or of composing fugues .. [] .. It's partly because of these [thinking] styles that creativity has an ambiguous relationship with freedom."*

Considering the many factors constituting the evaluation of what is deemed 'creative', raises core issues regarding how humans evaluate creativity; their aesthetic capacity and potentially that of other animals (e.g. as exhibited in, say, mate-selection). Galanter [18] suggests that perhaps the 'computational equivalent' of a bird or an insect (e.g. in evaluating mate selection) is all that is required for [computational] aesthetic evaluation:

> *"This provides some hope for those who would follow a psychological path to computational aesthetic evaluation, because creatures with simpler brains than man practice mate selection."*

In this context, as suggested in [16], the tastes of the individual in male bowerbirds are made visible when they gather collections of bones, glass, pebbles, shells, fruit, plastic and metal scraps from their environment, and arrange them to attract females [10]:

> *"They perform a mating dance within a specially prepared display court. The characteristics of an individual's dance or artefact display are specific to the species, but also to the capabilities and, apparently, the tastes of the individual."*

However the question of whether *'mate selection behaviour in animals implies making a judgement analogous to aesthetic judgement in humans'* is perhaps (pace Nagel's famous discussion 'What is it like to be a bat? '[25]) a fundamentally unanswerable question.

In contrast, the role of education (or training) in recognising 'good' and 'bad', 'creative' and 'non-creative' has been experimentally probed. A suggestive study investigating this topic by Watanabe [36] gathers a set of children's paintings, and then adult humans are asked to label the "good" from the "bad". Pigeons are then trained through operant conditioning to only peck at good paintings. After the training, when pigeons are exposed to a novel set of already judged children's paintings, they show their ability in the correct classification of the paintings.

This emphasises the role of learning training and raises the question on whether humans are fundamentally trained (or "biased") to distinguish good and/or creative work.

Another tightly related topic to swarm intelligence in this context is the creativity of social systems. Bown in [11] indicates that our creative capabilities are contingent on the objects and infrastructure available to us, which help us achieve individual goals, in two ways:

> *"One way to look at this is, as Clark does [13], in terms of the mind being extended to a distributed system with an embodied brain at the centre, and surrounded by various other tools, from digits to digital computers. Another way is to step away from the centrality of human brains altogether and consider social complexes as distributed systems involving more or less cognitive elements."*

Discussion on creativity and the conditions which make a particular work creative, have generated heated debate amongst scientists and philosophers for many years [27]; for a theoretical review on 'conditions of creativity'; the 'systems' view of creativity; cognitive approaches, etc. see also [32]. Although this article does not aim to resolve any of these issues (or even suggest that the presented work strongly fits and endorses the category of the 'computationally creative realm'), we investigate the performance of a swarm intelligence sketching system which, we suggest, highlights core issues inherent in exploring conceptual/artistic space(s).

3 CREATIVITY IN SWARMS

3.1 Freedom vs. Constraint

Both freedom and constraint have always been at the core of several definitions for creativity. Philip Johnson-Laird in his work on freedom and constraint in creativity [22] states:

> *"... for to be creative is to be free to choose among alternatives .. [] .. for which is not constrained is not creative."*

In swarm intelligence systems, the two phases of exploration and exploitation introduce the freedom and control the level of constraint. Pushing the swarms towards exploration, freedom is boosted; and by encouraging exploitation, constraint is more emphasised. Finding a balance between exploration and exploitation has been an important theoretical challenge in swarm intelligence research and over the years many hundreds of different approaches have been deployed by researchers in this field. In the presented work, two swarm intelligence algorithms are deployed: the algorithm which is responsible for the "intelligent" tracking of the line drawing is Particle Swarm Optimisation (PSO). This well-known algorithm, which mimics the behaviour of birds flocking, has an internal mechanism of balancing off the exploitation and exploration phases. However due to the weakness of the exploration in this algorithm, our system also deploys another nature inspired algorithm to overcome this weakness, Stochastic Diffusion Search (SDS), which mimics the behaviour of one species of ants (*Leptothorax acervorum*) foraging. Therefore, exploration is promoted by utilising the SDS algorithm, whose impact on different swarm intelligence algorithms has been scientifically reported using various measures and statistical analysis in several publications (e.g. [2, 3, 4, 5]) and the technical information on the integration of the two algorithms can be found in al-Rifaie et al. [2].

In the visualisation, the swarms are presented with a set of points (which constitute a line drawing – see Fig. 1) and are set to consider these points (one at a time) as their global optimum. In other words, the global optimum is dynamic, moving from one position to another and the swarms aim to converge over this dynamic optimum (Fig. 2).

As stated in the introduction, there have been several relevant attempts to create creative computer generated artwork using Artificial Intelligence, Artificial Life and Swarm Intelligence. Irrespective of whether the swarms are considered genuinely creative or not, their similar individualistic approach is not totally dissimilar to those of the "elephant artists" [37]:

> *"After I have handed the loaded paintbrush to [the elephants], they proceed to paint in their own distinctive style, with delicate strokes or broad ones, gently dabbing the bristles on the paper or with a sweeping flourish, vertical lines or arcs and loops, ponderously or rapidly and so on. No two artists have the same style."*

Similarly if the same line drawing (see Fig. 1) is repeatedly given to the swarms, the output sketches (e.g. Fig 2) made by the swarms, are never the same (see Fig. 4 to compare different sketches). In other words, even if the swarms process the same input several times, they will not make two identical sketches; furthermore, the outputs they produce are not merely randomised variants of the input. In order to demonstrate this claim qualitatively in an experiment, the output of the swarm-based system is compared against a simple randomised tracing algorithm, where each point in the line drawing could be surrounded with lines at a random distance and direction.

In Fig 3, only PSO algorithm is used to producing the sketch. This experiment is run in order to highlight the exploration (i.e. 'freedom') impact induced by SDS algorithm on the final sketch.

3.2 Swarmic Freedom versus Random Freedom

This part presents an experiment with the goal of contrasting the behaviour of the swarms to that of a group of random agents. In this experiment, the freedom of the swarm (i.e. *Swarmic Freedom*) is maintained by the swarm intelligence algorithms used in the system, whereas the freedom of the agents in the randomised algorithm is controlled by what we call the *Random Freedom*. These definitions are utilised here to highlight the potential of the swarms in exhibiting computational creativity.

The sketches in Fig. 5 (top and middle) show two outputs from a simple randomised algorithm when configured to exhibit limited 'random' variations in their behaviour (i.e. there is only small random distance and direction from the points of the original line drawing); comparing the two sketches, we note a lack of any significant difference between them. Furthermore, when more 'freedom' is granted to the randomised algorithm (by increasing the range in the

Figure 1. This figure shows a series of points that make a line drawing; sample line drawing after one of Picasso's sketches.

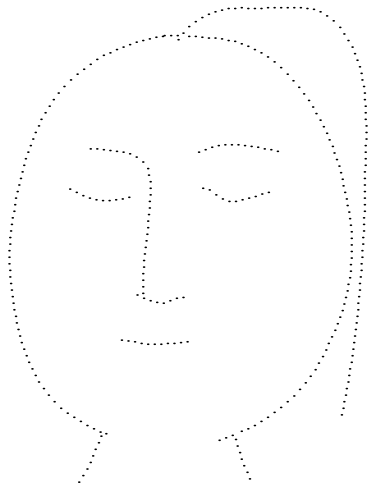
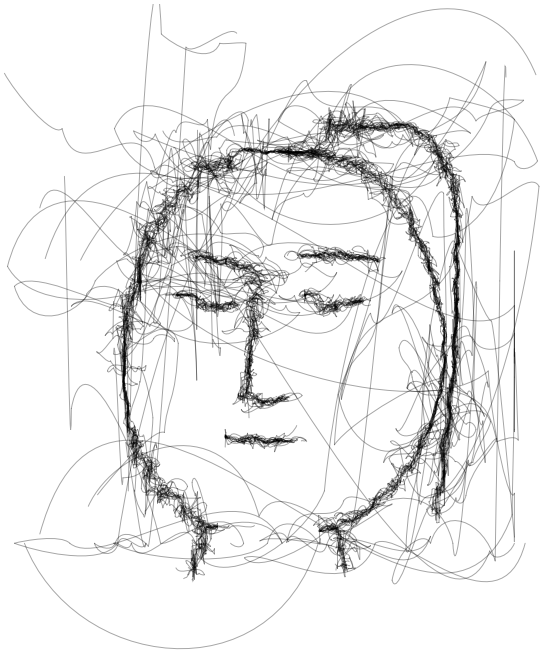


Figure 2. A sketch produced by the swarms.



underlying random number generator, which allows the technique to explore broader areas of the canvas), the algorithm soon begins to deviate excessively from the original line drawing. For this reason such randomisation results in a very poor - low fidelity - interpretation of the original line drawing (Fig. 5-bottom). In contrast, although the agents in the swarms are free to access any part of the canvas, the swarm-control mechanism (i.e. Swarm Freedom) naturally enables the system to maintain recognisable fidelity to the original input. In the randomised algorithm, contra the swarms system, it can be seen that simply by giving the agents more randomised behaviour (Random Freedom), they fails to produce more 'creative sketches'.

The Swarmic Freedom or 'controlled freedom' (or the 'tincture of madness') exhibited by the swarm algorithms (induced by the stochastic side of the algorithms) is crucial to the resultant work and is the reason why having the same line drawing does not result in the system producing identical sketches. This freedom emerges, among other influencing factors, from the stochasticity of SDS algorithm in picking agents for communication, as well as choosing agents to diffuse information; the tincture of madness in PSO algorithm is induced via its strategy of spreading the particles throughout the search space as well as the stochastic elements in deciding the next move of each particle.

In other words, the reason why the swarm sketches are different from the simple randomised sketches, is that the underlying PSO flocking component-algorithm constantly endeavours to accurately trace the input image whilst the SDS foraging component constantly endeavours to explore the wider canvas (i.e. together the two swarm mechanisms ensure high-level fidelity to the input without making an exact low-level copy of the original line drawing). Although the algorithms (PSO and SDS) are nature-inspired, we do not claim that the presented work is an accurate model of natural systems. Furthermore, whilst designing the algorithm there was no explicit 'Hundertwasser-like' attempt [26] by which we mean the stress on using curves instead of straight lines, as Hundertwasser considered straight lines not nature-like and tried not to use straight lines in his works to bias the style of the system's sketches.

3.3 Weak vs. Strong Computational Creativity

Before approaching the topic of weak or strong computational creativity, the difference between weak and strong AI is highlighted. In strong AI, the claim is that machines can think and have genuine understanding and other cognitive states (e.g. "suitably programmed machines will be capable of conscious thought" [12]); weak AI, in contrast, does not usually go beyond expecting the simulation of human intelligence. I.e. instantiating genuine "understanding" is not the primary concern in weak AI research.

An analogy could be drawn to computational creativity, extending the notion of weak AI to weak computational creativity, which does not go beyond exploring the simulation of human creativity; emphasising that genuine understanding is not the main issue in weak computational creativity. In strong computational creativity, the expectation is that the machine should be creative, have genuine understanding and other cognitive states as well as being capable of conscious thought.

Having a machine with conscious thought has provoked many critics, among whom John Searle made the most famous attack against strong AI in his Chinese Room argument [30]. Bishop [8] summarises Searle's Chinese Room Argument (CRA) as follows:

> "In 1977 Schank and Abelson published information [28] on a program they created, which could accept a simple story and then answer questions about it, using a large set of rules, heuristics and scripts. By script they referred to a detailed description of a stereotypical event unfolding through time. For example, a system dealing with restaurant stories would have a set of scripts about typical events that happen in a restaurant: entering the restaurant; choosing a table; ordering food; paying the bill, and so on. In the wake of this and similar work in computing labs around the world, some of the more excitable

Figure 3. A sketch produced by the swarms without SDS exploration.

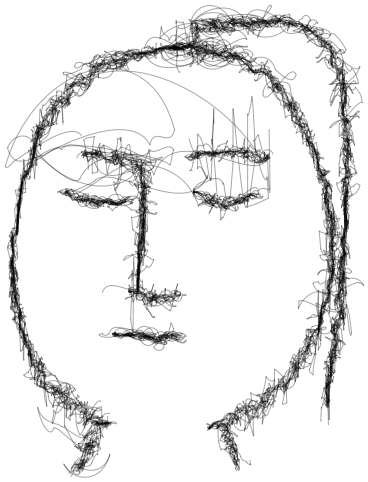
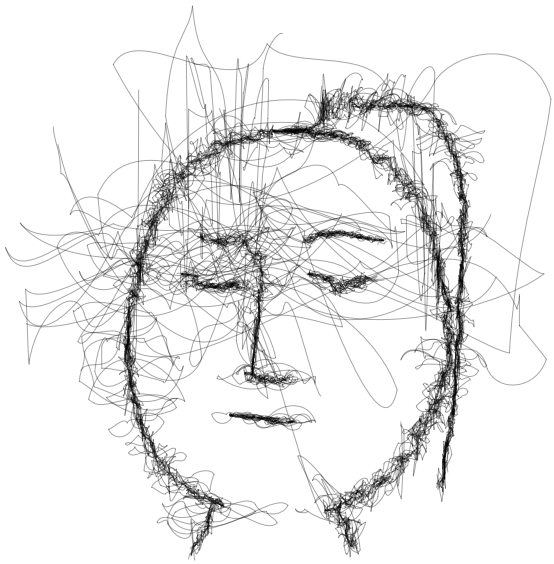


Figure 4. Different sketches of the swarms off a single line drawing.



proponents of artificial intelligence began to claim that such programs actually understood the stories they were given, and hence offered insight into human comprehension.

It was precisely an attempt to expose the flaws in the statements emerging from these proselytising AI-niks, and more generally to demonstrate the inadequacy of the Turing test[3], which led Searle to formulate the Chinese Room Argument.

The central claim of the CRA is that computations alone cannot in principle give rise to understanding, and that therefore computational theories of mind cannot fully explain human cognition. More formally, Searle stated that the CRA was an attempt to prove that syntax (rules for the correct formation of sentences:programs) is not sufficient for semantics (understanding). Combining this claim with those that programs are formal (syntactical), whereas minds have semantics, led Searle to conclude that 'programs are not minds'.

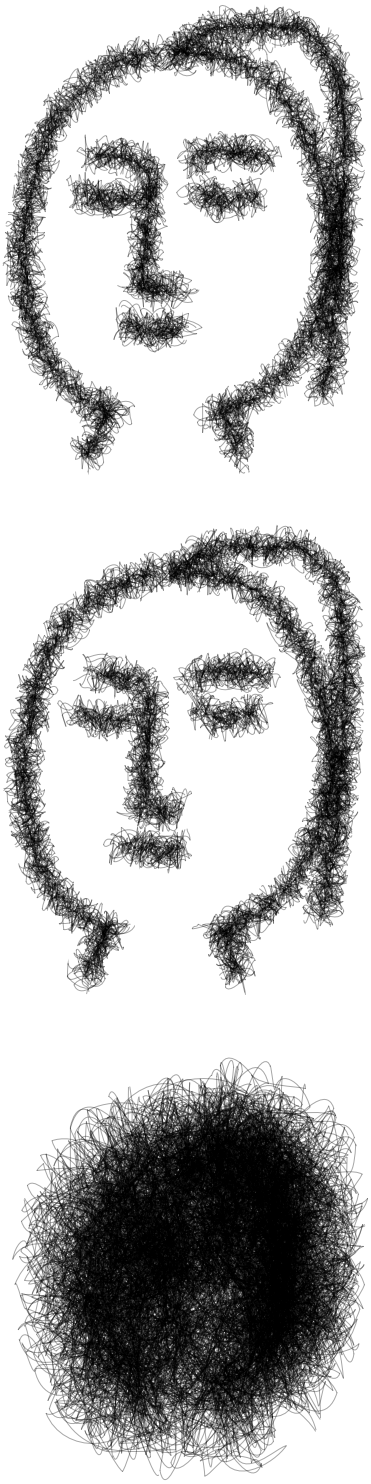
And yet it is clear that Searle believes that there is no barrier in principle to the notion that a machine can think and understand; indeed in MBP [Minds, Brains and Programs] Searle explicitly states, in answer to the question 'Can a machine think?', that 'the answer is, obviously, yes. We are precisely such machines'. Clearly Searle did not intend the CRA to target machine intelligence *per se*, but rather any form of artificial intelligence according to which a machine could have genuine mental states (e.g. understanding Chinese) purely in virtue of executing an appropriate series of computations: what Searle termed 'Strong AI'.

Searle argues that understanding, of say a Chinese story, can never arise purely as a result of following the procedures prescribed by any computer program, for Searle offers a first-person tale outlining how he could instantiate such a program, and act as the Central Processing Unit of a computer, produce correct internal and external state transitions, pass a Turing test for understanding Chinese, and yet still not understand a word of Chinese.

Searle describes a situation whereby he is locked in a room and presented with a large batch of papers covered with Chi-

[3] In what has become known as the 'standard interpretation' of the Turing test a human interrogator, interacting with two respondents via text alone, has to determine which of the responses is being generated by a suitably programmed computer and which is being generated by a human; if the interrogator cannot reliably do this then the computer is deemed to have 'passed' the Turing test.

Figure 5. The sketches of the swarms with random behaviour: This figure shows the sketches made with a simple randomised tracing algorithm, using random distance and direction from the lines of the original line drawing. The first two sketches (top and middle) use the same random distance (e.g. d) and the bottom sketch uses the random distance of $d \times 6$.



nese writing that he does not understand. Indeed, the monoglot Searle does not even recognise the symbols as being Chinese, as distinct from say Japanese or simply meaningless patterns. Later Searle is given a second batch of Chinese symbols, together with a set of rules (in English) that describe an effective method (algorithm) for correlating the second batch with the first, purely by their form or shape. Finally he is given a third batch of Chinese symbols together with another set of rules (in English) to enable him to correlate the third batch with the first two, and these rules instruct him how to return certain sets of shapes (Chinese symbols) in response to certain symbols given in the third batch.

Unknown to Searle, the people outside the room call the first batch of Chinese symbols 'the script', the second set 'the story', the third 'questions about the story' and the symbols he returns they call 'answers to the questions about the story'. The set of rules he is obeying they call 'the program'. To complicate matters further, the people outside the room also give Searle stories in English and ask him questions about these stories in English, to which he can reply in English.

After a while Searle gets so good at following the instructions, and the 'outsiders' get so good at supplying the rules he has to follow, that the answers he gives to the questions in Chinese symbols become indistinguishable from those a true Chinese person might give.

From an external point of view, the answers to the two sets of questions, one in English the other in Chinese, are equally good; Searle, in the Chinese room, have passed the Turing test. Yet in the Chinese language case, Searle behaves 'like a computer' and does not understand either the questions he is given or the answers he returns, whereas in the English case, *ex hypothesi*, he does. Searle contrasts the claim posed by some members of the AI community - that any machine capable of following such instructions can genuinely understand the story, the questions and answers - with his own continuing inability to understand a word of Chinese; for Searle the Chinese symbols forever remain ungrounded."

We suggest that Searle's famous thought experiment similarly targets the notion of 'strong computational creativity'. I.e. Searle using a similar "room" could gets so good at following the rules that the strings of symbols he outputs from the room successfully control a 'Strong computer creative art' system producing works judged to have artistic merit by people outside the room; even though Searle-in-the-room remains ignorant of art and art practise. Hence, until the challenge of the Chinese room has been fully met, the authors urge caution in predicating 'strong' notions of creativity to any computational system.

4 CONCLUSION

In this paper, we have discussed the potential of the swarms in exhibiting 'weak computational creativity'. This specific work described herein uses swarm intelligence techniques to explore the difference between using Random Freedom and Swarmic Freedom in the visualisation of the swarms 'tracing' line drawings; this work highlights the features of swarm-regulated difference versus simple-random difference in the production of such 'sketches' by computer. We stressed on the significant impact of both freedom and constraint on the emergent creativity, and presented a discussion on how these two concepts are mapped onto exploration and exploitation, the two most infamous phases in the swarm intelligence world. The so described computational artist is the result of merging two swarm intelligence algorithms (SDS and PSO), preserving freedom (exploration) and constraint (exploitation) respectively.

5 CODA

Leit-motif: Although we distance ourselves from claims of strong computational creativity, in faint homage to Turing's Imitation Game and Harre & Wang's physical implementation of the Chinese room experiment [20], we asked Chiara Puntil, a human artist from Goldsmiths, to adopt the 'style' of the swarms and to produce some sketches (Fig 6) based on the 'style' of the line drawing in Fig. 2.

ACKNOWLEDGEMENT

The authors would like to thank the contributing artist, Chiara Puntil, an alumna from Goldsmiths College, Department of Arts.

REFERENCES

[1] Mohammad Majid al-Rifaie, Mark Bishop, and Ahmed Aber, 'Creative or not? birds and ants draw with muscles', in *AISB 2011: Computing and Philosophy*, pp. 23–30, University of York, York, U.K., (2011). ISBN: 978-1-908187-03-1.

[2] Mohammad Majid al-Rifaie, Mark Bishop, and Tim Blackwell, 'An investigation into the merger of stochastic diffusion search and particle swarm optimisation', in *GECCO '11: Proceedings of the 2011 GECCO conference companion on Genetic and evolutionary computation*, pp. 37–44, Dublin, Ireland, (2011). ACM.

[3] Mohammad Majid al-Rifaie, Mark Bishop, and Tim Blackwell, 'An investigation into the use of swarm intelligence for an evolutionary algorithm optimisation', *International Conference on Evolutionary Computation Theory and Application (ECTA 2011)*, (2011).

[4] Mohammad Majid al-Rifaie, Mark Bishop, and Tim Blackwell, 'Resource allocation and dispensation impact of stochastic diffusion search on differential evolution algorithm; in', in *Nature Inspired Cooperative Strategies for Optimisation (NICSO 2011), Springer*, (2011).

[5] Mohammad Majid al-Rifaie, Mark Bishop, and Tim Blackwell, 'Information sharing impact of stochastic diffusion search on differential evolution algorithm', in *Journal of Memetic Computing: Nature Inspired Cooperative Strategies for Optimization*, eds., David Pelta and et al. Springer Berlin Heidelberg, (2012). submitted.

[6] Mohammad Majid al-Rifaie, Mark Bishop, and Suzanne Caines, 'Creativity and autonomy in swarm intelligence systems', in *Cognitive Computation: Computational Creativity, Intelligence and Autonomy*, eds., Mark Bishop and Yasemin Erden. Springer, (2012). DOI: 10.1007_s12559-012-9130-y.

[7] S. Aupetit, V. Bordeau, N. Monmarche, M. Slimane, and G. Venturini, 'Interactive evolution of ant paintings', in *The 2003 Congress on Evolutionary Computation, 2003. CEC'03.*, volume 2, pp. 1376–1383, (2004).

[8] M. Bishop, 'A view inside the chinese room', *The Philosopher*, 28(4), 47–51, (2004).

[9] M.A. Boden, *Creativity and Art: Three Roads to Surprise*, Oxford University Press, 2010.

[10] Gerald Borgia, 'Complex male display and female choice in the spotted bowerbird: specialized functions for different bower decorations', *Animal Behaviour*, 49, 1291–1301, (1995).

[11] O. Bown, 'Generative and adaptive creativity', in *In Computers and Creativity*, eds., Jon McCormack and Mark d'Inverno. Berlin: Springer, (2011).

[12] Rob Callan, *Artificial Intelligence*, Palgrave Macmillan, 2003.

[13] A. Clark, *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*, Oxford University Press, 2003.

[14] J. Collomosse and P. Hall, 'Genetic paint: A search for salient paintings', *Applications of Evolutionary Computing*, 437–447, (2005).

[15] C. J Curtis, 'Loose and sketchy animation', in *ACM SIGGRAPH 98 Electronic art and animation catalog*, p. 145, (1998).

[16] A Dorin and K. Korb, 'Creativity refined. in computers and creativity', in *In Computers and Creativity*, eds., Jon McCormack and Mark d'Inverno. Berlin: Springer, (2011).

[17] A. Etzioni, A. Ben-Barak, S. Peron, and A. Durandy, 'Ataxia-telangiectasia in twins presenting as autosomal recessive hyper-immunoglobulin m syndrome', *IMAJ*, 9(5), 406, (2007).

[18] P. Galanter, 'Computational aesthetic evaluation: Past and future', in *In Computers and Creativity*, eds., Jon McCormack and Mark d'Inverno. Berlin: Springer, (2011).

[19] G. Greenfield, 'Evolutionary methods for ant colony paintings', *APPLICATIONS OF EVOLUTIONARY COMPUTING, PROCEEDINGS*, 3449, 478–487, (2005).

[20] R. Harre and H.T. Wang, 'Setting up a real 'chinese room: an empirical replication of a famous thought experiment1', *Journal of Experimental & Theoretical Artificial Intelligence*, 11(2), 153–154, (1999).

[21] A. Hertzmann, 'Paint by relaxation', in *Computer Graphics International 2001. Proceedings*, pp. 47–54. IEEE, (2001).

[22] P. N Johnson-Laird, 'Freedom and constraint in creativity', in *The nature of creativity: contemporary psychological perspectives*, ed., Robert J. Sternberg, p. 202219. Cambridge University Press, (1988).

[23] N. Monmarche, S. Aupetit, V. Bordeau, M. Slimane, and G Venturini, 'Interactive evolution of ant paintings', in *2003 Congress on Evolutionary Computation*, ed., B. McKay et al, volume 2, pp. 1376–1383. IEEE Press, (2003).

[24] L. Moura and V. Ramos, 'Swarm paintings–nonhuman art', *ARCHITOPIA book, art, architecture and science*, 5–24, (2007).

[25] T. Nagel, 'What is it like to be a bat?', *The Philosophical Review*, 83(4), 435–450, (1974).

[26] P. Restany, *Hundertwasser: the painter-king with the five skins: the power of art*, Taschen America Llc, 2001.

[27] A. Rothenberg and C.R. Hausman, *The creativity question*, Duke University Press Books, 1976.

[28] R.C. Schank, R.P. Abelson, et al., *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*, volume 2, Lawrence Erlbaum Associates Hillsdale, NJ, 1977.

[29] S. Schlechtweg, T. Germer, and T. Strothotte, 'Renderbots–multi-agent systems for direct image generation', in *Computer Graphics Forum*, volume 24, pp. 137–148, (2005).

[30] J. Searle, 'Minds, brains, and programs', *Behavioral and Brain Sciences*, 3(3), 417–457, (1980).

[31] Y. Semet, U. M O'Reilly, and F. Durand, 'An interactive artificial ant approach to non-photorealistic rendering', in *Genetic and Evolutionary Computation–GECCO 2004*, pp. 188–200, (2004).

[32] R.J. Sternberg, *The nature of creativity: Contemporary psychological perspectives*, Cambridge Univ Pr, 1988.

[33] G. Turk and D. Banks, 'Image-guided streamline placement', in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pp. 453–460. ACM, (1996).

[34] P. Urbano, 'Playing in the pheromone playground: Experiences in swarm painting', *Applications on Evolutionary Computing*, 527–532, (2005).

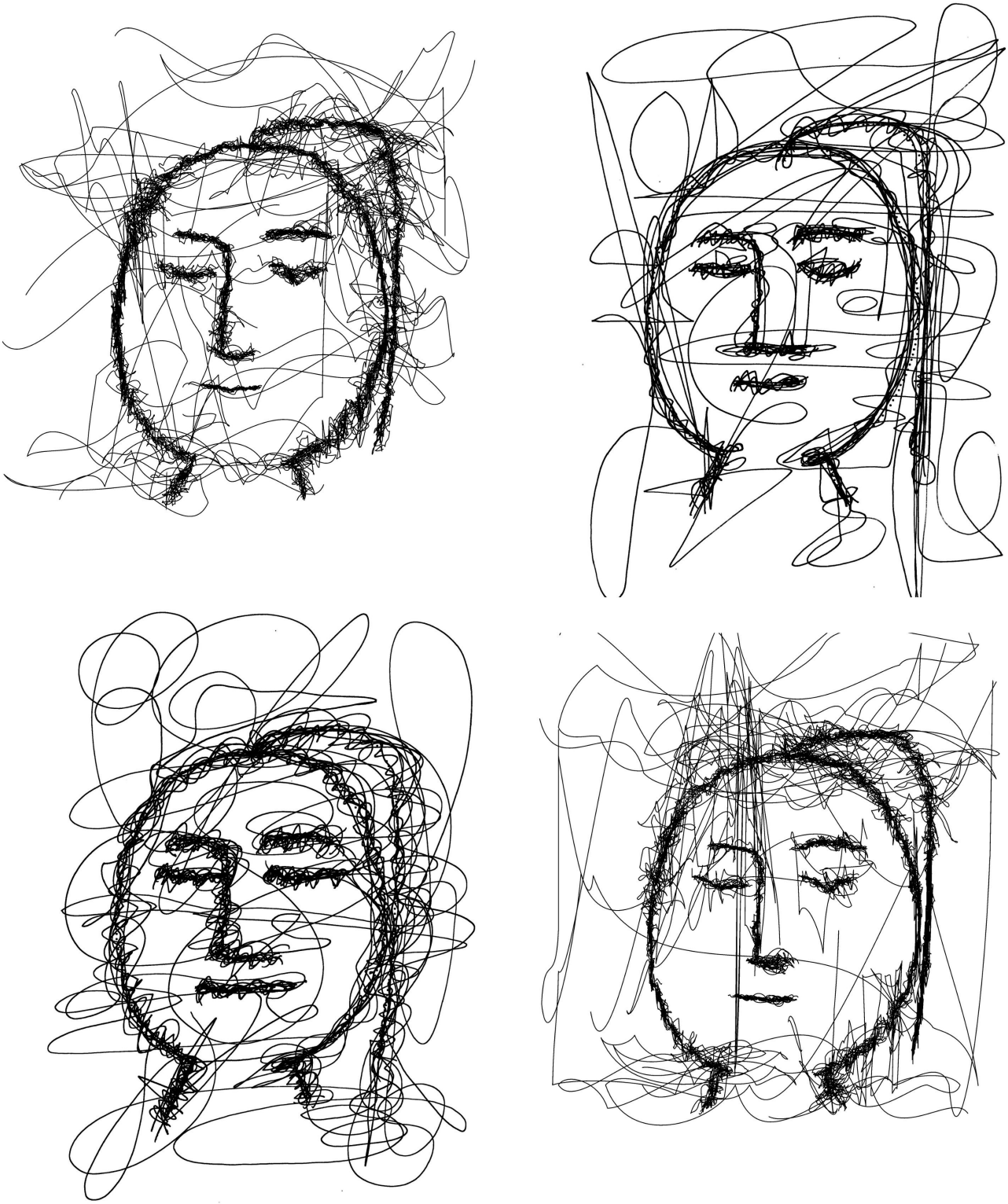
[35] P. Urbano, 'Consensual paintings', *Applications of Evolutionary Computing*, 622–632, (2006).

[36] Shigeru Watanabe, 'Pigeons can discriminate "good" and "bad" paintings by children', *Animal Cognition*, 13(1), (2009).

[37] Aum-Mon Weesatchanam, *Are Paintings by Elephants Really Art?*, The Elephant Art Gallery, 31 July 2006.

[38] M. Zhao and S.C. Zhu, 'Customizing painterly rendering styles using stroke processes', in *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Non-Photorealistic Animation and Rendering*, pp. 137–146. ACM, (2011).

Figure 6. Two of the sketches are produced by the swarms and two are made by a human artist.



Mathematical Models of Desire, Need and Attention

Alexander J Ovsich

Abstract. Desire plays an important role in the explanation of behavior in general, for example, in the contemporary Belief-Desire theories. These theories (for example, Bratman's Belief-Desire-Intention theory) are widely used in the AI applications. However, there is neither much literature, nor even consensus about the meaning and definition of desire. There is not much clarity about the concepts and mechanisms of need and attention either.

The author presents here simple, closely linked mathematical models of desire, need, and attention. They are based upon the hedonistic principle proclaiming that animals and humans alike are driven by striving to maximize pleasantness of their internal state (Pleasantness of the State of this Subject("PSS"). What directly follows from this principle is that for such a subject (S), the most important characteristic of any phenomenon (X) should be how much X influences the process of maximization, how much X increases or decreases PSS, that is measured by the magnitude and direction of its change (Δ PSS).

I propose that terms such as 'desire,' 'want,' and their cognates describe PSS change associated with (caused by) a phenomenon: $DESIRE_{S,X} = \Delta PSS_{S,X}$;

if $\Delta PSS_{S,X} > 0$, then X is called desirable;

if $\Delta PSS_{S,X} < 0$, then X is called undesirable**.**

The magnitude of the PSS change is what is called "strength of desire":

STRENGTH of the $DESIRE_{S,X} = |DESIRE_{S,X}| = |\Delta PSS_{S,X}|$.

Need is defined here as a term describing a periodic or cyclical desire.

There is another direct inference from the hedonistic principle: the more a phenomenon affects the process of PSS maximization, i.e. the larger a PSS change it creates or the stronger a desire is associated with it, the more attention should a subject pay to it:

$$ATTENTION_{S,X} \sim |\Delta PSS_{S,X}| \sim |DESIRE_{S,X}|;$$

$$ATTENTION_{S,X} = k\,|DESIRE_{S,X}|.$$

Considering that an overall attention of a subject S at any given moment t ($ATTtotal_t$) is distributed between a number of objects (1 to n) and that it has an upper limit ($ATTmax_{S,t}$):

$$ATTmax_{S,t} >= ATTtotal_{S,t} =$$

$$k|DESIRE_{S,t,1}| + k|DESIRE_{S,t,2}| + \ldots + k|DESIRE_{S,t,n}|.$$

Boston College, USA, email: ovsich@bc.edu

Keywords: Attention; Desire; Hedonism; Need; Pleasure; Pleasantness; Want.

1 LACK OF CLARITY AND CONSENSUS

There are many 'ways' [1] and 'faces' [2] of desire, but there is one fundamental question about the meaning and definition of desire that is the focus of this paper. There is neither much literature nor consensus about the notion of desire. Schueler [3], who "... focused on contemporary philosophers..." noted that "... the views I am criticizing suffer from a deep ambiguity in terms such as 'desire', 'want' and their cognates". Almost a decade later Frankfurt [4] called the notion of desire "rampantly ubiquitous" and wrote:

> Moreover, its various meanings are rarely distinguished; nor is there much effort to clarify how they are related. These matters are generally left carelessly undefined in the blunt usages of common sense and ordinary speech.

The level of ambiguity in understanding desire is such that the validity of the notion of desire itself is sometimes questioned or even denied outright. For example, DeLancey [5] wrote:

> Since my concern in this book is with basic emotions and other motivational states, I will on several occasions discuss the inappropriateness of the philosopher's notion of desire; it is hard to overestimate the harm that this notion has done to moral psychology, action theory, and other aspects of philosophy of mind.
> … (for example, there are many kinds of motivational states, but no generic one corresponding to the philosophical notion of desire)… .

However, as Marks [1, p. 10] carefully noted:

> …it may well be the case, as I believe, that there remains a single, significant, psychological phenomenon appropriately named "desire." If so, then it is this – desire proper – which, ultimately, constitutes the subject matter of the theory of desire.

His belief is shared by the author of this paper.

2 HEDONISTIC APPROACH TO DESIRE

Schroeder [2, pp. 27-31] identified two main types of the desire theories – motivational and hedonistic; he considered hedonistic theory to be superior to the motivational[1]. Indeed, the hedonistic approach to desire has a very long and impressive history. Aristotle [7] directly defined desire through pleasure: "Everything, too, is pleasant for which we have the desire within us, since desire is the craving for pleasure" and the same can be said about Spinoza [8]. As formulated by Mill [9] "…desiring a thing and finding it pleasant, aversion to it and thinking of it as painful, are phenomena entirely inseparable or, rather, two parts of the same phenomenon." Schroeder [2, p.27], referring to this Mill's opinion, wrote "Mill is not the only distinguished historical figure to have considered such a view." Schroeder further elaborated: "Hobbes, Hume, and Kant apparently had similar thoughts, though interpretation of these thinkers is difficult" [2, p. 185].

In line with his clearly hedonistic definition of desire as "the craving for pleasure" quoted above, Aristotle [7, 2.2, 1378a 31-3,] not only defines anger as a desire for revenge [10], [11] or retaliation [12], but also provides rather detailed descriptions of what it means at the hedonic level [7, 1.11, 1371a; also see 2.2, 1378b]:

> Revenge, too, is pleasant; it is pleasant to get anything that it is painful to fail to get, and angry people suffer extreme pain when they fail to get their revenge; but they enjoy the prospect of getting it.

It is important here to note that desire for revenge (anger) involves a positive hedonic change, transition from the hedonically negative to the hedonically positive state experienced even while imagining 'the prospect of getting it'.

Aristotle's hedonistic approach to desire was echoed by Locke who defined desire as follows: "The uneasiness a man finds in himself upon the absence of anything whose present enjoyment carries the idea of delight with it, is that we call desire" [13]. Desire for Locke is also about the hedonic gap between the more negative hedonic level ("uneasiness") of the state of the desiring subject without an object of desire and the more positive hedonic level ("enjoyment") with it. As for Aristotle, Locke's interpretation of desire is also about the positive hedonic change associated with the desired phenomenon.

The vital fact of the matter here is that such a hedonic gap, a positive hedonic change associated with the object of desire is a regular property of the subjective experience of desire. This is true for the "low" physiological desires as well as for the "high" psychological desires. This sameness allows one to express desire for an action, power or sex metaphorically as being "action or power hungry," "hungry for the loved one."

[1] He also added "the third face of desire" - his own "reward and punishment" theory of desire that was sharply criticized - see, for example, review of Katz [6].

3 FORMULAS OF DESIRE AND ITS STRENGTH

The model of desire presented in Ovsich [15], [16], [17] and discussed here is based upon the Hedonistic Principle declaring that animals and humans alike are driven by a striving to maximize pleasantness of their internal state (Pleasantness of the State of a Subject or PSS[2] here). The direct inference from the Hedonistic Principle is that (one of) the most important characteristics of any phenomenon for a subject driven to maximize PSS is how much this phenomenon maximizes (or minimizes) PSS. For the human subject it should also mean that words and expressions describing PSS changes ought to be notable and widely used.

Ovsich proposed that terms such as 'desire,' 'want,' and their cognates describe PSS change (ΔPSS) associated with (caused by) a phenomenon:

1. expressions calling a phenomenon X 'desirable', 'wanted'
   etc., for the subject S characterize X as a factor of maximization of PSS for the subject; that these expressions associate X with the positive $\Delta PSS_{S,X}$;
2. X associated with negative $\Delta PSS_{S,X}$ is called 'undesirable', 'unwanted';
3. X associated with zero $\Delta PSS_{S,X}$ is called indifferent, though sometimes it is called undesirable in the sense of the lack of desire.

The common feature in cases two and three is a non-positive (zero or negative) change of PSS ($\Delta PSS_{S,X} <= 0$) or an absence of the positive change of PSS. It indicates, that,

- a subject reports a presence or absence of desire for a phenomenon depending upon the presence or absence of the positive change of PSS associated with that phenomenon;
- what is usually called 'desire' of X is a positive change of PSS associated with X;
- an object of desire is a factor of PSS maximization.

From the hedonistic viewpoint it is quite clear why a positive rather than a negative or zero change of PSS is used as the bases for terms 'desire' and 'want' describing PSS alteration. According to the Hedonistic Principle, a subject is looking for maximization of PSS that is represented by a positive PSS change, $\Delta PSS_{S,X} > 0$. The use of the negative prefix to describe something as 'Undesirable', 'Unwanted', points to the opposite (negative) to the positive PSS change that subjects are seeking or to the absence of the positive PSS change.

[2] Pleasantness/valence is a complex variable. Emotions and a number of sensations, possess their own pleasantness of specific modality. All these P/U can be experienced at the same time and are represented by a complex structure, that changes at every given moment. We call it here a Pleasantness of the State of a Subject (PSS. PSS is quite close to what is called a Valence of the Core Affect in [17, [18]. For more details see [16].

If we interpret desire as an algebraic variable that can be positive or negative (where the 'desirable X' means that X is an object of the positive desire and 'Undesirable X' means that X is an object of the negative desire), then we can define the desire of X in general as a term describing a change of PSS ($\Delta PSS_{S,X}$) associated with X. Here is the definition of a desire: *a subject's (S) desire for X is a word to describe a change of the Pleasantness of the State of this Subject* ($\Delta PSS_{S,X}$) *associated with (or 'caused' by) the perception or imagination of X. Desirability of X for S is an ability of X to maximize/minimize PSS that is characterized by* $\Delta PSS_{S,X}$.

Below is the formula of desire that incorporates all three types of the ΔPSS, and where S is a subject experiencing desire, X is an object of desire, ΔPSS is the change of the Pleasantness of the State of the Subject:

$$DESIRE_{S,X} = \Delta PSS_{S,X} \qquad (1)$$

The above definition and formula of desire are consistent both with hedonistic/utilitarian approach to desire and with the contemporary point of view, that "…the primary linkage of the notion of desire to a notion other than itself is to the notion of affect – pleasure or displeasure in the widest sense" [19].

A desire is often characterized or measured by its strength. Both positive and negative desire can be experienced as strong or weak. This means that the strength of desire is a sign-independent characteristic of desire. Therefore, a mathematical sign of the magnitude or an absolute value (|value|) should be applied to express strength of the subject's (S) desire for X ($\Delta PSS_{S,X}$):

$$\text{Strength of S desire for X} = |DESIRE_{S,X}| = |\Delta PSS_{S,X}| \qquad (2)$$

Experimental support of this model of desire is demonstrated in [16].

4 NEED AS A PERIODIC/CYCLICAL DESIRE

Experiencing a need means feeling the corresponding desire. As Audi [20] wrote, "Human needs are innate and quickly give rise to desires". S. L. Rubinshteòin [21] has declared that *desire is a concrete form of the need's existence*[3]. If a subject experiences a desire for X repeatedly or regularly it is usually said that the subject *needs* X. This is clearly demonstrated by the needs that emerge and cease to exist with age or during changing conditions, for example, the needs for sex, smoking, or drugs. The origination/disappearance of such needs is acknowledged when the corresponding desire begins/ stops being regular or repeated. *Need is a term used for a periodic or cyclical desire*. This is true for all kinds of need including any need for food, sex, activities, drugs, etc.. A need is characterized by the strength and frequency of its desire.

Need, being a cyclical process is like a 'wave' of desire. All needs have definable features. Dissatisfaction of any need of a subject negatively affects PSS, and this decline of PSS grows with time. At the same time, P of perceived or imagined objects of this need's satisfaction for the subject goes up.

These two aspects are easily recognizable in the following description of Bertrand Russell [22],

> …it seems clear that what, with us, sets a behavior-cycle in motion is some sensation of the sort which we call disagreeable. Take the case of hunger : we have first an uncomfortable feeling inside, producing a disinclination to sit still, a sensitiveness to savory smells, and an attraction towards any food that there may be in our neighborhood.

This means that the hedonic gap between PSS without the object(s) of a need satisfaction and PSS with it grows. *This gap is a desire and its magnitude is its strength.*

Satisfaction of any need of a subject produces exactly opposite effects: PSS grows as a result of satisfaction of a need and P of the objects of this need's satisfaction goes down. As the hedonic gap of desire gets smaller, desire gets weaker all the way down to the satiation point when ΔPSS of desire becomes equal to zero – desire is satisfied, and then disappears. At this time, the opposite side of the desire cycle starts again.

5 ATTENTION AND HEDONISM

Another direct inference from the Hedonistic Principle is that the more a phenomenon influences the process of PSS maximization the more attention should be paid to it. The effect of X on the process of PSS maximization is measured by the magnitude of the PSS change ($|\Delta PSS_{S,X}|$) associated with X, that according to the above model of desire is the

$$\text{Strength of S Desire for X} = |DESIRE_{S,X}| = |\Delta PSS_{S,X}|.$$

In the first approximation, attention of a subject S toward a phenomenon X can be considered to be simply proportional to the strength of desire for it:

$$ATT_{S,X} = k|(\Delta PSS_{S,X}| = k|DESIRE_{S,X}|, \qquad (3)$$

where k is a positive coefficient of proportionality.

The model of attention to a 'single' phenomenon above is a sheer abstraction, because in reality a subject always perceives multiple phenomena. This model, however, represents an approximation of a real situation, where the subject concentrates mainly on one phenomenon in the center of attention. The higher the percentage of total attention paid to the phenomenon in the center of attention, the closer this model comes to reality.

There are some situations when a phenomenon is singled out and placed in the center of attention. This occurs in a process of choice making when the elements of choice are appraised by a subject and attitudes toward them are formed one by one, until a 'new' phenomenon catches the attention of a subject and is appraised or perhaps an 'old' phenomenon is re-appraised. This also happens when a phenomenon becomes 'the chosen one' and is placed in the center of attention, while all competing phenomena are pushed to the periphery of attention. At this early stage of this analysis, all but the one 'central' phenomenon will be disregarded.

[3] Translated by Ovsich.

6 ATTENTION TO A SINGLE PHENOMENON

Let's analyze the formula of attention (3) to see if it describes different situations correctly.

Case #1. $\Delta PSS_{S,X} > 0$ or $DESIRE_{S,X} > 0$
In this the case X is a factor of PSS *maximization*, meaning that the subject wants X.

IF $\Delta PSS_{S,X} > 0$, $DESIRE_{S,X} > 0$ THEN $ATT_{S,X} > 0$

According to the formula (3), $ATT_{S,X}$ increases/decreases if the positive desire $DESIRE_{S,X}$ increases/decreases. The greater the desire for X by a subject the more attention is paid thereto.

Case #2. $\Delta PSS_{S,X} < 0$ or $DESIRE_{S,X} < 0$
In this case, X is a factor of PSS *minimization,* meaning that the subject does not want X.

If $\Delta PSS_{S,X} < 0$, $DESIRE_{S,X} < 0$ then $ATT_{S,X} > 0$

The formula $ATT_{S,X} = k|DESIRE_{S,X}|$ illustrates that the stronger the negative desire for X (the more bothersome or undesirable X is) the more attention is paid to it.

Cases #1 and #2 show that according to the formula (3) a subject pays attention to both desirable and undesirable phenomena. The more desirable or undesirable it is – that is to say, the greater the strength of the (+) or (-) desire for the phenomenon, the more attention will be paid to it.

The substance of this matter is that *eliminating the sources of PSS minimization* is just as important for the hedonistic process as *acquiring the sources of PSS maximization* because of the integrative character of PSS. Adding $100 to an account affects its balance in the same way as canceling a $100 debt. A subject's concentration on the sources of a positive ΔPSS for their *exploitation* as well as concentration on the sources of a negative ΔPSS for their *elimination* are equally important for this process of PSS maximization. Attention paid to X doesn't depend on the sign of ΔPSS_X or a desire for X but only on the magnitude of the PSS change that is the strength of desire for x. In summary, attention paid to X is *sign-independent* of whether X is desirable or undesirable, but depends only on the strength of desirability/undesirability of X.

Case #3. $\Delta PSS_{S,X} = 0$ or $DESIRE_{S,X} = 0$

If $\Delta PSS_{S,X} = 0$, $DESIRE_{S,X} = 0$ then $ATT_{S,X} = 0$

If X is indifferent to a subject (meaning that X doesn't affect the PSS maximization of a subject, that there is no + or - desire for X) then a subject won't pay any attention to X. No attention at all is paid to the hedonically indifferent phenomena.

A graph for attention as a function of desire is a vertical "V" with its point at the zero of the crossing of the horizontal axis of desire and the vertical axis of attention.

7 HEDONISTIC RESOLUTION OF THE FRAME PROBLEM

Though Case #3 above is the least important hedonically, it is the most important statistically. At any given moment, animals, including humans, do not pay attention to the great majority of phenomena accessible to them because they are indifferent to them. This allows them to concentrate on the small percentage of phenomena that are important for their existence and well-being. Zero desire experienced toward indifferent phenomena that require no attention is a powerful filter and eliminator affording great protection for the limited resources of a small creature facing an endless Universe. This is the essence of "… the human talent for *ignoring* what should be ignored, while staying alert to relevant recalcitrance when it occurs" [23].

I would suggest that imitation of this mechanism and the mechanism of hedonic orientation in general is key to the resolution of one of the fundamental problems of Artificial Intelligence, called "the qualification problem" by McCarthy [24], usually called a "frame problem", and described by Dennett [23, p. 161] as follows:

> What is needed is a system that genuinely *ignores* most of what it knows, and operates with a well-chosen portion of its knowledge at any moment. Well chosen, but not chosen by exhaustive consideration. How, though, can you give a system *rules* for ignoring - or better, since explicit rule-following is not the problem, how can you design a system that reliably ignores what it ought to ignore under a wide variety of different circumstances in a complex action environment?

I agree with McFarland's point of view [25]:

> It is worth noting that animals do not suffer from the frame problem, and this may be because they have a *value system* (see Chapter 8), the cost and risks involved in their decision-making acting as constraints on their behavior.

The above analysis of the formula (3) for attention shows that this formula gives an accurate basic description of some fundamental features of attention. It correctly illustrates the fact that both positive and negative influences on a subject's PSS get attention, and that the degree of attention to a phenomenon is proportional to the magnitude of its desirability. It is fair to say that at least in some measure this formula applies.

8 ATTENTION TO MULTIPLE PHENOMENA

In reality, a subject is always simultaneously perceiving multiple phenomena, because the fact of the matter is that at any given moment the attention of a subject is distributed between a multitude of simultaneously perceived phenomena[4]. I propose that the total volume of attention of a subject S perceiving n phenomena at the moment t ($ATTtotal_{S,t}$) can be described as the sum of attention paid to

[4] See, for example, Damasio [26].

each of them:

$$ATTtotal_{s,t} = ATT_{s,t,1} + ATT_{s,t,2} + ... + ATT_{s,t,n} \quad (4)$$

Now, let's merge it (4) with formula for attention to a single phenomenon (3) by replacing every component of the right part of (4), representing attention to one of the n phenomena, with its expression from (3):

$$ATTtotal_{s,t} = k|\Delta PSS_{s,t,1}| + k|\Delta PSS_{s,t,2}| + \ldots + k|\Delta PSS_{s,t,n}| =$$

$$= k|DESIRE_{s,t,1}| + k|DESIRE_{s,t,2}| + \ldots + k|DESIRE_{s,t,n}| \quad (5)$$

This formula (5) clearly demonstrates that attention is distributed between n simultaneously perceived phenomena unevenly, in accordance with the magnitude of their desirability [5].

9 CENTER OF ATTENTION

Attention has its periphery and its most focused or 'brightest' area which is usually called the 'center of attention'. Let's assign numbers to perceived phenomena in descending order from 1 to n, in accordance with the volume of attention paid by a subject to each of them:

$$ATT_{s,t,1} > ATT_{s,t,2} > \ldots > ATT_{s,t,n}$$

Thus, the number one ($ATT_{s,t,\mathbf{1}}$) will be assigned from now on to the phenomenon having the most attention or being at the center of attention. According to the formula (5), this indicates the phenomenon with the largest positive or negative influence on PSS change |ΔPSS| - the one that is most desirable or undesirable, i.e. corresponding to the strongest desire:

$$ATT_{s,t,1} > ATT_{s,t,2} > ... > ATT_{s,t,n}$$

or

$$|\Delta PSS_{s,t,1}| > |\Delta PSS_{s,t,2}| > \ldots > |\Delta PSS_{s,t,n}|$$

or

$$|DESIRE_{s,t,1}| > |DESIRE_{s,t,2}| > \ldots > |DESIRE_{s,t,n}|$$

10 GENERAL FORMULA OF ATTENTION

There is one more general feature of attention that has to be taken in consideration: attention has an upper limit. In the words of Csikszentmihalyi [27]:

> The main assumption I shall be making is that attention is a form of a psychic energy needed to control the stream of consciousness, and that attention is a limited psychic resource (p. 337).

This means that at any moment (t) there is a maximum or an upper limit for the attention of a subject S ($ATTmax_{s,t}$) and that at any moment t this maximum is not less than the total attention of a subject:

$$ATTmax_{s,t} >= ATTtotal_{s,t,1\text{-}n} = \quad (6)$$

$$= k|\Delta PSS_{s,t,1}| + k|\Delta PSS_{s,t,2}| + \ldots + k|\Delta PSS_{s,t,n}| =$$

$$= k|DESIRE_{s,t,1}| + k|DESIRE_{s,t,2}| + \ldots + k|DESIRE_{s,t,n}|$$

It is important, that the general formula of attention (6) includes within itself the formula for attention to a single phenomenon (3) as a particular case corresponding to the situation when n equals to 1:

$$ATTmax_{s,t} >= ATTtotal_{s,t,1} = k|\Delta PSS_{s,t,1}| = k|DESIRE_{s,t,1}|$$

There are the following variables in the general formula of attention:

1. a subject S;
2. t – time;
3. $ATTmax_{s,t}$ (maximum of attention of S available at t);
4. $ATTtotal_{s,t,1to\ n}$ (total disbursed attention at t);
5. $|DESIRE_{s,t,n}|$ (strength of desire of S for n at t);
6. n - number of the phenomena perceived by S simultaneously at the moment t.

Let's find out how this formula works with different combinations of values for these variables/parameters and how the formula's implications reflect reality.

11 UPPER LIMIT OF ATTENTION

According to the formula (6), if the left part of equation becomes smaller, then the right part has to be lessened as well too. It can be reduced by the number (n) of phenomena that are paid attention to, and/or by a decrease of the magnitude of their desirability for the subject:

if $ATTmax_{s,t} \rightarrow 0$

then $ATTtotal_{s,t} \rightarrow 0$;

and

$$(k|DESIRE_{s,t,1}| + \ldots + k|DESIRE_{s,t,n}) \rightarrow 0$$

It can happen because:

$n \rightarrow 0$

and/or

$|DESIRE_{s,t,1}|, \ldots. |DESIRE_{s,t,n}) \rightarrow 0$

This corresponds to what can be observed in reality. $ATTmax_{s,t}$ represents the upper limit of attention of a subject S available at the moment t. If it grows, a subject is able to pay even more attention to the same number n of perceived phenomena or can increase their number. Conversely, if $ATTmax_{s,t}$ is diminished, then a subject ought to pay less attention to the same number (n) of phenomena and/or has to decrease their number.

$ATTmax_{s,t}$ goes down when a subject gets tired. For example, with the subject getting more and more fatigued,

[5] I suggest that in the first approximation k is the same for all the simultaneous objects of attention from 1 to n.

desirability of the current activities and attention paid to them decrease. One loses the desire to do anything. The only desire that remains at this point is to do nothing, to get rest, to pay no further attention to anything at all.

12 CHANGE OF DESIRABILITY AND ATTENTION REDISTRIBUTION

Here we will consider what happens with distribution of attention if desirability of one of the n simultaneously perceived phenomena changes.

Case #1, Change of *positive* desirability of the phenomenon X; $DESIRE_{s,x} > 0$.

If any positive value gets larger, then its absolute value or magnitude is also enlarged. So, if positive desire grows, then its magnitude or strength ($|DESIRE_{s,x}|$) also gets larger. According to the formula (3)

$$ATT_{s,x} = k|DESIRE_{s,x}|$$

attention towards the phenomenon grows together with the strength of the desire for it or with its desirability.

With the additional attention paid to one of the n phenomena, that particular one will move up in the 'attention hierarchy'; it will earn an attention 'promotion'. This phenomenon would change its place in the row of the decreasing attention levels corresponding to n different phenomena perceived at the same time t.

$$ATT_{s,t,1} > ATT_{s,t,2} > ... > ATT_{s,t,n}$$

Its position will move from right to left in the above formula and its number placement (from 1 to n) will decrease until it becomes the number one phenomenon in the center of attention. The reverse process, an attention 'demotion' can be said to occur according to this formula when the strength of desirability of the phenomenon diminishes.

Attention 'promotion' and 'demotion' as prescribed by this formula does take place in reality. A good illustration of such a promotion is provided by taking note of a growing desire corresponding to an ongoing unsatisfied need. Such a desire strengthens until it gets into the center of attention of a subject together with those objects and ways of its satisfaction. This situation has been analyzed from a different point of view in the prior discussion of need.

In the course of satisfaction of a need the reverse process takes place. Desire gets weaker, and the attention paid to the objects and actions of satisfaction for this desire decreases, and as such, these objects and acts move out from the center of the subject's attention to its periphery and finally completely out of range. The center of attention gets overtaken by other phenomena.

Case #2. Change of *negative* desirability (undesirability) of the phenomenon X: $DESIRE_{s,x} < 0$.

If any negative value gets more negative, then its absolute value or magnitude is getting larger. So, if negative desire grows, if its object gets more undesirable, then the magnitude or strength of its undesirability ($|DESIRE_x|$) gets larger. The formula (3) shows that attention towards the phenomenon grows together with the strength or magnitude of its *undesirability*.

As in the case #1, with the additional attention paid to one of the n phenomena, that particular one will move up in the 'attention hierarchy', will earn an attention 'promotion'. This phenomenon would change its place in the row of the decreasing attention levels corresponding to n different phenomena perceived at the same time t.

$$ATT_{s,t,1} > ATT_{s,t,2} > ... > ATT_{s,t,n}$$

Its position will move from right to left in the above formula and its number placement (from 1 to n) will decrease until it becomes the number one phenomenon in the center of attention. The reverse process, an attention 'demotion' can be said to occur according to this formula when the strength of *undesirability* of the phenomenon diminishes.

A good illustration of the cases where attention grows toward *undesirables* is provided by any kind of the increase of discomfort or unpleasantness, for example strengthening of toothache or hunger pangs. The more unpleasant and undesirable something becomes for a subject, the more attention is drawn thereto. The less unpleasant and undesirable it becomes due to the action of a painkiller or food intake, the less attention is paid thereto.

Comment about cases #1 and #2.

The similarity in changes of attention in the above cases one and two illustrate the independence of attention paid to a phenomenon from the positive or negative value sign of its desirability. It is also interesting that the dissatisfaction of a need can serve as an example for both cases. An object of a need's satisfaction, as well as corresponding subjective state both get an attention promotion that escalates during the time of the ongoing need dissatisfaction. An object of need (for example, food) rises in the attention hierarchy through an *increase* in the desirability of this object while the specific subjective state of the dissatisfaction of that need (hunger, thirst, etc.) gets an attention promotion through the *decrease* in the desirability for that specific state.

In these cases, nature uses *both* of its major tools of orientation - positive and negative in order to drive a subject to satisfy a need. It pushes a subject *away from* the subjective state of dissatisfaction of a need and simultaneously *pulls toward* the object or way of its satisfaction. It makes the current state of the dissatisfied subject unpleasant and thus undesirable while at the same time, making the objects of satisfaction that much more desirable.

13 HEDONIC "PRICING" AND REDISTRIBUTION OF ATTENTION

According to the Hedonistic Principle, animals and humans alike are driven by hedonic striving to maximize their PSS. Therefore, a major tool of their orientation is their hedonic 'pricing' through attaching a factor of Pleasantness/Unpleasantness to a phenomenon in order to establish it as positive or negative factor of PSS maximization and determine its desirability. By using variants of reward and punishment, like the carrot and stick scenario, both nature and

society affix hedonic sticker-prices of what is pleasant or unpleasant and set values on good and bad. Adjustment of this P or hedonic 'pricing' is a most significant instrument in the alteration of animal and human orientation and choice. This adjustment has been experimentally studied by Cabanac [28], [29], who called it "alliesthesia" [28, p.1105]:

> In order to avoid using a whole sentence saying that a given external stimulus can be perceived either as pleasant or unpleasant depending upon signals coming from inside the body, it may be useful to use a single word to describe this phenomenon. I hereby propose the word alliesthesia (8) coming from esthesia (meaning sensation) and allios (meaning changed).

Let us suppose that a subject perceives the same n phenomena for a given time when $ATTmax_{s,t}=ATTtotal_{s,t}$, but attention that is required for one of n phenomena grows.

$$ATTmax_{s,t} = ATTtotal_{s,t} = ATT_{s,t,1}+ATT_{s,t,2}+...+ ATT_{s,t,n}$$
$$k|DESIRE_{s,t,1}| + k|DESIRE_{s,t,2}| +\dots + k|DESIRE_{s,t,n}|.$$

This formula shows that as one of the n phenomena (number x <= n) gathers more attention, then the other (n-1) phenomena will have less attention left to them. If maximum of available attention ($ATTmax_{s,t}$) is not used up ($ATTmax_{s,t}<ATTtotal_{s,t}$), then the total disbursed attention ($ATTtotal_{s,t}$) can be increased up to the level of $ATTmax_{s,t}$. Now is the time for a subject to become more alert. Conversely, if maximum of available attention is already used up ($ATTmax_{s,t} = ATTtotal_{s,t}$), then the total of available attention (ATT_{tot}) must be redistributed. If the remainder of attention is not enough for the rest (n-1) of the evident phenomena, then some of them will receive no attention at all. Hence, a reduction of the number (n) of the perceived phenomena takes place. At this point, attention becomes more focused or narrowed. If attention to x grows so great that it requires all of the available attention of a subject, then all of it has to be spent on X only:

$$ATTmax_{s,t} = ATTtotal_{s,t} = ATTmax_{s,t,x}$$

It may be that an adult deeply concentrated on inner thoughts or a child running after a ball may not pay enough attention to that oncoming car. The more concentrated a subject is on something, the more difficult it will be for anything else to catch one's attention. And conversely, if the concentration of attention for a subject is low, then any new phenomena can easily get to the center of attention. For example, a bored child in the classroom is just looking for anything new to switch attention to.

A good example of the narrowing down of attention is the case where a basic need of a subject has not been satisfied for a long period of time. (A 'long' period of time can here be probably defined as a multiple of the regular or average period of time between satisfactions of this need). In this case, objects and images of the subject's need become more and more desirable and demand more and more attention. They gradually push everything out of the center of the subject's attention to the periphery until they have completely taken over. Eventually the objects and images of the subject's need become 'super-values' for that moment. Think toilet.

This converges with one of the basic postulates of Ethology, as described by Cabanac [30], because the strongest desire corresponds to the 'most urgent need' of this postulate: "One basic postulate of Ethology is that behavior tends to satisfy the most urgent need of the behaving subject (Tinbergen, 1950; Baerends, 1956)".

14 CHANGE OF THE OBJECTS OF ATTENTION

In reality, a subject constantly perceives new phenomena. An important distinctive quality of new phenomena is the unpredictability of their appearance. At any moment, new phenomenon can appear and make demands on a subject's attention. The following redistribution of attention, possible promotion of a very hedonically important phenomenon to the center of attention can be as sudden as its appearance. The stage of the attention distribution described by the equation

$$ATTmax_{s,t} = ATTtotal_{s,t} = ATTmax_{s,t,x}$$

can be reached at once in case of an unforeseen extreme danger or excitement.

For example, while walking down the street one perceives numerous objects but pays little attention to most of them. A subject can see many cars on the street and pay them no attention at all. But the distribution of a subject's attention changes right away with the recognition of a friend inside a car, or when it seems that one of these cars is going to hit the subject.

15 COMPUTER MODELING OF DESIRE, NEED AND ATTENTION

Formalization of any process in clear mathematical terms makes it possible to create its computer model. I believe it can happen with the proposed hedonistic models of desire, need and attention. Together they represent a considerable part of the choice mechanism of the autonomous, hedonistically driven system. Let's call such a system a "*hedonicus"*.

One of the advantages of the computer/robotic implementation of a *hedonicus* is the similarity of its design with some features of its initial creator – *Homo hedonicus.* This similarity should offer ease of the *hedonicus-to-hedonicus* communication because they will speak the same language.

16 CONCLUSION

This article presents closely linked mathematical models of desire, need and attention. They are simple, intuitive and, to the best knowledge of the author, are the only hedonistic/ mathematical models of desire, need and attention available. The author believes that they can be accommodated in the design of autonomous systems.

REFERENCES

[1] Marks, J. (Ed., 1986). *The Ways of Desire: New Essays in Philosophical Psychology on the Concept of Wanting*. Chicago: Precedent Publishing.

[2] Schroeder, T. (2004). *Three Faces of Desire*. Oxford: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195172379.001.0001

[3] Schueler, G. F. (1995). *Desire. Its Role in Practical Reason and the Explanation of Action*. Cambridge, Massachussetts: MIT Press, p. 6.

[4] Frankfurt, H. G. (2004). *The reasons of love.* Princeton, N.J.: Princeton University Press, p. 10.

[5] DeLancey, C. (2002). *Passionate Engines:What Emotions Reveal About Mind and Artificial Intelligence*. Oxford: Oxford University Press, p. ix.

[6] Katz, L. D. (2005). Three Faces of Desire. *Philosophical Reviews, University of Notre Dame*. http://ndpr.nd.edu/review.cfm?id=3861

[7] Aristotle (2004). *Rhetoric* (T. W. Rhys, Trans.): Dover Publications, I, 11, 1370b

[8] Spinoza, B.(1674/1955). *On the Improvement of the Understanding. The Ethics,.Correspondence.* New York: Dover Publications, Inc., Proposition XXXVII.

[9] Mill, J. S. (1861/1961). *Utilitarianism*. New York: Priest, O. Macmillan, p. 49.

[10] Konstan, D. (2006). *The Emotions of the Ancient Greeks: Studies in Aristotle and Classical Literature (Robson Classical Lectures)*. Toronto: University of Toronto Press, p. 41.

[11] Kenny, A. (1963). *Action, Emotion, and Will.* London: Routledge and Kegan Paul, p 193.

[12] Tailor C. C. W. (1986). Emotions and Wants. In J. Marks (Ed.) *The Ways of Desire: New Essays in Philosophical Psychology on the Concept of Wanting* (pp. 217-231). Chicago: Precedent Publishing, p.231.

[13] Locke, J. (1690/1824). *An Essay Concerning Human Understanding.* Edit 12, Vol. 1. London: Baldwin, Printer, Book II, Chapter 20, Section 6.

[14] Ovsich, A. J. (1998a). Outlines of the Theory of Choice: Attitude, Desire, Attention, Will. *Proceedings of the 1998 IEEE International Symposium on Intelligent Control (ISIC) Held Jointly with the International Symposium on Computational Intelligence in Robotics and Automation and the Intelligent Systems and Semiotics (ISAS), pp. 503-510*

[15] Ovsich, A. J. (1998b). Outlines of the Theory of Choice: Attitude, Desire, Attention, Will. *Proceedings of the 1998 Twentieth World Congress of Philosophy.* http://www.bu.edu/wcp/Papers/Acti/ActiOvsi.htm

[16] Ovsich, A., Cabanac. M. (2012). Experimental Support of the Mathematical Model of Desire. *International Journal of Psychological Studies. Vol.4. No. 1.* doi:10.5539/ijps.v4n1p66 http://dx.doi.org/10.5539/ijps.v4n1p66

[17] Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110, 145–172. http://dx.doi.org/10.1037/0033-295X.110.1.145

[18] Barrett, L. F. (2006). Valence is a basic building block of emotional life. *Journal of Research in Personality,* 40, 35-55. http://dx.doi.org/10.1016/j.jrp.2005.08.006

[19] Strawson, G. (2010). *Mental Reality*. The MIT Press, p.284.

[20] Audi, R. (1993). *Action, Intention, and Reason*. Ithaca, New York: Cornell University Press, p. 29.

[21] Rubinshteòin, Sergeòi Leonidovich. (1957). *Bytie i soznanie; o meste psikhicheskogo vo vseobshcheòi vzaimosvþiÆazi þiÆavleniòi materialñnogo mira.* Moskva: Izd-vo Akademii nauk SSSR.

[22] Russell, B. (1921). *The Analysis of Mind*. London: G. Allen & Unwin, ltd, p. 67.

[23] Dennett, D. C. (1990). Cognitive Wheels: The Frame Problem of AI _In Boden M. E. (Ed.). *The philosophy of artificial intelligence* (pp. 147-171). Oxford: Oxford University Press, p.162.

[24] McCarthy, J. (1968). Programs with Common Sense Proceedings of the Teddington Conference on the Mechanization of Thought Processes, London.Repr. In M.

[25] McFarland, D. (2008). *Guilty Robots, Happy Dogs: TheQuestion of Alien Minds*. Oxford: Oxford University Press, p.156.

[26] Damasio A. R. (1994). *Descartes' Error. Emotion, Reason, and the Human Brain,* Bard, an Avon Book, p.199.

[27] Csikszentmihalyi, M. (1978). Attention and the Holistic Approach to Behavior. In Kenneth S. Pope and Jerome L. Singer. (Eds.). *The Stream of consciousness: scientific investigations into the flow of human experience* (pp. 335-358). New York: Plenum Press, 1978, p.337..

[28] Cabanac, M. (1971). Physiological role of pleasure. *Science, 173*, 1103-1107.

[29] Cabanac, M. (2010). *The Fifth Influence. The Dialectics of Pleasure.* iUniverse Books, 2010, ISBN: 978-1-4401-8836-7. English translation of the: *La cinquième influence, ou La dialectique du plaisir*. 2003. Québec: Presses de l'Université Laval.

[30] Cabanac, M. (2000). Pleasure, the prerational intelligence. In H. Ritter, H. Cruse & J. Dean (Eds.), *Prerational Intelligence: Adaptive Behavior and Intelligent Systems Without Symbols and Logic* (pp. 201-213). Dordrecht: Kluwer Academis Publishers, p.1.