Proceedings of the Symposium

2nd Symposium on Computing and Philosophy

A symposium at the AISB 2009 Convention (6-9 April 2009)
Heriot-Watt University, Edinburgh, Scotland

Symposium Chairs
Mark Bishop

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

http://www.aisb.org.uk/

ISBN - 1902956826

2nd Symposium on Computing and Philosophy

A one-day symposium at AISB 2009 (6-9 April 2009).
http://www.doc.gold.ac.uk/seminars/AISB09/Philosophy+Computing.html

PROGRAMME CHAIRS

Dr Mark Bishop, Goldsmiths College London, UK

INTRODUCTION

The convergence of computing and philosophy has a lineage going back to Leibniz but it is not until the work of Alan Turing and the appearance of electronic computers in the mid-20th century that we arrive at a practical intersection between computing and philosophy. Precursors to the theories and programs of interest to this AISB Symposium on Computing and Philosophy include: the Turing Test as outlined in Turing's seminal reflection on thinking machines; the AI work of Herb Simon and Alan Newell with the Logic Theorist; Rosenblatt's Perceptron - a biologically inspired pattern matching device and Grey Walter's Turtle - an early example of embodied Cybernetic Artificial Intelligence (A.I).

The aim of this symposium is to advance the philosophical study of computing in general by exploring the philosophical analysis of central concepts in computer science, the application of computational principles to traditional philosophical problems and computational modelling of philosophical assumptions. To this end the group of topics selected for discussion in the symposium include: foundational issues of cognition; distributed cognition; embodiment; inference; computational platforms for embodied thought experiments; embodied mathematics; algorithms and the Turing test; machine ethics.

On behalf of the organising committee of this second AISB Computing And Philosophy symposium, I would like to thank, for their support in both organising the event and in refereeing the submissions to the symposium, all the members of the program committee. Lastly I hope that all authors and participants find the day enjoyable and this second symposium worthwhile.

TOPICS

Topics of interest include but are not limited to:

- Constructivism; enactivism; second order cybernetics
- Dynamic systems theories of cognition
- Sensorimotor theories of perception
- Artificial life; computer modelling in biology; simulation of behaviour
- Machine understanding; Searle's Chinese room argument; the Turing test
- Biosemiotics
- Embodied A.I.; robotics
- Virtual reality; computer-mediated communication
- Philosophy of information / technology
- Information and computer ethics
- Metaphysics (distributed processing, emergent properties, formal ontology, network structures, etc.)

PROGRAMME COMMITTEE

Table of Contents

Do socio-technical systems cognise?

Olle Blomberg [1]

Abstract. The view that an agent's cognitive processes sometimes include proper parts found outside the skin and skull of the agent is gaining increasing acceptance in philosophy of mind. One main empirical touchstone for this so-called *active externalism* is Edwin Hutchins' theory of *distributed cognition* (DCog). However, the connection between DCog and active externalism is far from clear. While active externalism is one component of DCog, the theory also incorporates other related claims, which active externalists may not want to take on board. DCog implies a shift away from an organism-centred cognitive science to a focus on larger socio-technical-cum-cognitive systems. In arguing for this shift, proponents of DCog seem to accept that socio-cultural systems have some form of agency apart from the agencies of the individuals inside them. I will tentatively suggest a way in which such a notion of agency can be cashed out.

1 Introduction

In "The Extended Mind" [8], Andy Clark and David Chalmers ask where the mind stops and the rest of the world begins. They argue that bits of our environment sometimes become proper parts of our cognitive processes. In other words, contrary to the received view in cognitive science, cognitive processes sometimes loop out beyond the skin and skull. This is a claim about the location and boundaries of cognition.

A closely related issue is what the "unit of analysis" should be in cognitive science. The unit of analysis is the system or set of interactions that needs to be analysed in order to reach a correct understanding of how organisms cognize and behave. Presumably, if cognitive processes extend beyond the skin and skull, so should the unit of analysis. But an extended unit of analysis may be recommended on less radical grounds too, as Robert Rupert has pointed out [27]. It is enough to claim that cognition is deeply embedded in the world — without looping out into it — in order to conclude that "we can properly understand the traditional subject's cognitive processes only by taking into account how the agent exploits the surrounding environment to carry out her cognitive work" [27, p. 395].[2] While the boundaries of the unit of analysis and the boundaries of cognition are not necessarily the same then, they are clearly connected.

The *distributed cognition* approach (henceforth DCog) is probably the approach in cognitive science that has widened the unit of analysis the most. In DCog, socio-technical systems (made up of socially organized individuals equipped with tools and technologies) are treated as cognitive systems. [3] For example, the cognitive anthropologist Edwin Hutchins, in what is arguably the canonical account of DCog [17], provides a detailed ethnography of a navigation team steering a large military vessel into port. He analyses the navigation team as a cognitive system in which mental-cum-cultural representations are created, transformed and propagated.

While Hutchins' work is one of the main empirical touchstones of the philosophical extended mind movement, the relation between DCog and the philosophy that has drawn on it is unclear. This is partly due to the promiscuous use of the 'distributed cognition' label as more or less synonymous with 'the extended mind', 'active externalism', 'vehicle externalism', 'locational externalism' etc. [8, 15, 31]. In this paper, I will use Clark and Chalmers' label *active externalism* to refer to these philosophical positions collectively. Clark and Chalmers [8] refer to Hutchins' research as an example of empirical work that "reflects" their active externalism. Other active externalists (such as Susan Hurley and Robert Wilson) as well as "active internalists" (such as Fred Adams, Ken Aizawa and Robert Rupert) also refer to DCog as a sort of empirical counterpart of active externalism [15, 31, 2, 27].

There is a clear focus on "socially distributed cognition" in the DCog literature. This is a phenomenon that is largely absent in discussions about active externalism. Clark and Chalmers' [8] mention the possibility of socially *extended* cognition, where one thinker's mental state is partly constituted by the state of another thinker, but in such a case, the cognitive system is still firmly centred on the brain of an individual human organism. However, the socio-technical systems that are typically studied using the DCog framework are *not* centred on an individual organism. DCog thus departs from Clark's "organism-centered" [7, p. 139] active externalism, in the sense that there is often no clear locus of control which can be attributed to any one organism (but not in the sense that there may be no organisms involved at all).[4]

I will not enter the debate between active externalists and active internalists. My aim is rather to clarify what the relation is between (Hutchins' version of) DCog and active externalism.[5] In the next section, I present the DCog approach and tease out four theoretical-philosophical claims that make up the approach's theoretical backbone. One of these claims is tantamount to a commitment to active externalism. In the following sections, I consider whether some of the arguments that have been used to support active externalism can

[1] University of Edinburgh, United Kingdom, email: K.J.O.Blomberg@sms.ed.ac.uk

[2] Rupert also makes use of the concept 'unit of analysis', but perhaps in a slightly different way. He characterises active externalists as claiming that "the unit of analysis should be the organism and certain aspects of its environment treated together, as a single, unified system." [27, p. 395] My concept 'unit of analysis' is intended to be separate from 'cognitive system' so that even an active internalist can claim that "the unit of analysis should be the organism and certain aspects of its environment", although she would reject that they in the end should be treated "together, as a single unified system."

[3] Such systems are sometimes also referred to as 'socio-cultural systems' or 'distributed cognitive systems' in the DCog literature.

[4] However, Christine Halverson, a former student of Hutchins, states that "*DCog focuses on the socio-technical system*, which usually (but not necessarily) includes individuals." [11]

[5] Rupert [27, pp. 391–2, 425n59] also briefly discusses the relation between DCog and active externalism.

be used to support a widening of the unit of analysis to cover socio-technical systems. I argue that this is doubtful and, considered as theory of human cognition, DCog seems to rest on a contentious claim about socio-technical systems having a form of agency.

2 The distributed cognition approach

DCog grew out of ethnographic studies of people interacting with each other and with various tools in organisational settings. Such socio-technical systems are conceptualised through the theoretical lens of DCog as both computational and cognitive. In Hutchins' analysis of naval navigation, the activity of the navigation team is described using the symbol-shuffling framework of traditional cognitive science, but applied to a unit of analysis that includes not only several mariners, but also various representational artifacts.[6]

To give some flavour of research informed by DCog, I here provide a brief summary of Hutchins' analysis of a type of navigation activity. When Hutchins did his fieldwork, a navy ship that was near land and in restricted waters had to have its position plotted on the chart (map) at intervals of a few minutes. In such situations, a team of about five people had to be involved in "the fix cycle". To fix a ship's position, two lines of sight from the ship to known visual landmarks have to be drawn on the chart (the ship should be where the lines intersect on the chart). Simplifying slightly, the fix cycle ran as follows: with the help of special telescopic sighting devices called alidades, two "bearing takers" determined the bearing (direction) of one landmark each; they reported the bearings over a telephone circuit to a "bearing timer-recorder" who jotted them down in the bearing log; the "plotter", standing beside the bearing time-recorder, then plotted the lines of sight on the chart to determine the ship's position.

Hutchins glosses this fix cycle in a computational framework drawn from traditional cognitive science:

> The task of the navigation team [...] is to propagate information about the directional relationships between the ship and known landmarks across a set of technological systems until it is represented on the chart. Between the situation of the ship in the world and the plotted position on the chart lies a bridge of technological devices. Each device (alidade, phone circuit, bearing log etc.) supports a representational state, and each state is a transformation of the previous one. Each transformation is a trivial task for the person who performs it, but, placed in the proper order, these trivial transformations constitute the computation of the ship's position. [16, pp. 206–7]

From a DCog perspective, the members of the navigation team together with their tools and social organisation make up a *cognitive system* that keeps the ship on track. It seems appropriate to think of the navigation activity as instantiating a form of computation, but why think of the distributed computational process as a *cognitive* process? Are all computational processes cognitive? Or just those that are somehow hooked up in the right way to a biological organism? I do not question that it may be fruitful to conceptualise and study a socio-technical system as computational systems for various reasons. But is it fruitful for cognitive science to adopt the socio-technical system as a unit of analysis? Will this increase our understanding of the nature and manifestation of human *cognition*? Why should these systems be studied by *cognitive science* rather than, say, social science?

Later on, I will engage with these questions. But first, I present four distinct theoretical-philosophical claims that are part of DCog and relate them to the current philosophical debate about active externalism.

2.1 Into the wild

Empirical research informed by DCog has primarily been descriptive and based on ethnographic observation. According to Hutchins, much research on cognition "in the lab" (arguably a highly atypical setting for human cognition) is based on unexamined assumptions about what a human mind is for. One of the supposed pay-offs of ethnographic studies of cognition "in the wild" is to expose these assumptions and provide "a refinement of a functional specification for the human cognitive system" [17, p. 371]. According to Hutchins, cognitive science needs to get a richer empirically grounded conception of its explananda.

Hutchins calls the approach he recommends *cognitive ethnography*. A cognitive ethnography is a description of a "cognitive task world" of some specific setting. Hutchins claims that we in fact know very little about such everyday cognitive task worlds since "our folk and professional models of cognitive performance do not match what appears when cognition in the wild is examined carefully." [17, p. 371] One systematic such mismatch that cognitive science is suffering from, according to Hutchins, consists in mistaking the cognitive properties of socio-technical systems for cognitive properties of individuals considered in isolation [17, p. 355]. Hutchins argues that recognition of this mistake should lead one to suspect that the performance of cognitive tasks such as navigation "requires internal representation of much less of the environment than traditional cognitive science would have led us to expect." [17, p. 132]

In sum, DCog incorporates a methodological commitment that I call the ETHNOGRAPHY claim:

> **ETHNOGRAPHY**: Cognitive science is operating with an inadequate functional specification of the mind. Ethnographic descriptions of cognitive activities in the wild can provide a better specification for cognitive science in the lab to work with.

Note that this claim in itself does not touch on the issue of where cognitive processes are to be found, it merely points out there is a gap in our knowledge about the range, variety, and constitution of everyday activities in which cognitive processes are somehow involved.

2.2 Computation in socio-technical systems

While DCog departs from traditional cognitive science in many ways, its core, the computational model of mind, is retained. Computation is broadly conceived as the "creation, transformation, and propagation of representational states" so that it can be applied both to what happens inside and outside the heads of individuals [17, pp. xvi, 49]. Hutchins actually argues that while the notion of computation as symbol manipulation was metaphorically applied to the individual mind (in the head), it is a *literal* description of what occurs within (some?) socio-technical systems [17, pp. 363–4].

[6] Other settings studied under the auspice of DCog include the cockpit of a commercial airliner [18], a telephone hotline group [1], software programming teamwork [9], a neuroscience laboratory [3], work practice in an engineering company [26], and trauma resuscitation teamwork [29] In Hutchins' terminology, these are all cases that exemplify social distribution of cognition. While Hutchins usually presents DCog as a theory about the nature of human cognition [17, 20], it should be noted that DCog is also used as an analytical framework in Human-Computer Interaction (HCI) and in Computer-Supported Cooperative Work (CSCW) [33, 14, 11]. Interestingly, Clark [7, p. 96] actually takes HCI to be a field that house "nascent forms" of a science of the extended mind.

> [T]he computation observed in the activity of the larger system can be described in the way that cognition had been traditionally described that is, as computation realized through the creation, transformation, and propagation of representational states. [17, p. 49]

It is a bit unclear whether Hutchins believes DCog to be a theory of socio-technical systems in general or only of a symbol-shuffling subset of them. In [17, p. 363] and [19, p. 67] Hutchins sometimes suggests that it is a framework restricted for describing a subset of systems, but in later writings DCog "refers to a perspective on all of cognition, rather than a particular kind of cognition" [14, p. 3] (see also [20, p. 376]).[7]

We thus get the COMPUTATION claim:

> **COMPUTATION**: (i) A socio-technical system is a computational system, in which "representational states are created, transformed and propagated", and (ii) cognitive science should take it as a unit of analysis.

I take COMPUTATION to constitute the core of DCog. Its first part, (i), sets DCog apart from other socio-cultural approaches to cognition, while the second part (ii) sets it apart from traditional internalist cognitive science. Note that the claim in (ii) is not that cognitive science should *exclusively* take socio-technical systems as its unit of analysis.

2.3 Crossing old boundaries

Hutchins typically does not merely construe socio-technical systems as computational systems, but also as cognitive systems.[8] In calling socio-technical systems "cognitive", Hutchins seems to accept *something like* Clark and Chalmers' Parity Principle.[9] It is the functional-computational contributions of a process that makes it cognitive, not whether it occurs on one side or the other of a skin or skull boundary. In an article co-authored with James Hollan and David Kirsh, he writes:

> Distributed cognition looks for cognitive processes, wherever they may occur, on the basis of the functional relationships of elements that participate together in the process. A process is not cognitive simply because it happens in a brain, nor is a process noncognitive simply because it happens in the interactions among many brains. [14, p. 175]

This in itself need not imply that the boundaries of the cognition of individuals need to be redrawn. One can imagine several brain-bound cognitive agents interacting in such a way, with each other and with their tools, that they collectively make up a larger cognitive system (so that there are several brainbound cognitive systems nested within a larger one). However, Hutchins argues that this is the wrong picture. One important advantage of having a single framework for describing both what goes on inside and outside the heads of individuals, Hutchins argues, is that this highlights that "the normally assumed boundaries of the individual are not the boundaries of the unit described by *steep gradients in the density of interaction among media*." [17, p. 157, my emphasis] He claims he has "developed a language of description of cognitive events that is unaffected by movement across old boundaries." [19, p. 65]

As I interpret Hutchins, the location of these steep gradients determines where the boundaries of cognitive systems. The criterion can be used to analyse the relevant boundaries of socio-technical systems, as well as individuals working with tools. For example, in his analysis of how a bearing taker finds a specific landmark to read and report the bearing, the system is not restricted by the bearing taker's biological boundaries. Instead, it includes, at one point, the degree scale and the tick hairline presented to the bearing taker as he aligns the alidade with the landmark, and it then shifts as activity progresses: "The active functional system thus changes as the task changes. A sequence of tasks will involve a sequence of functional systems, each composed of a set of representational media." [17, p. 157]

Hutchins' criterion for determining system boundaries is similar to John Haugeland's proposed *bandwidth criterion* for deciding whether the mind is a distinct component in a brain-body-world system [12]. Following Herbert Simon [30], Haugeland suggests that systems should be decomposed according the pathways and bandwidth of information flow in the systems. A system is made up of components that interact with each other over interfaces. Interfaces are points of well-defined low-bandwidth interaction between components. Components are made up of parts that interact at much higher bandwidth and in ill-defined ways (relative to the interaction that is mediated by the components interfaces). If a system can be analysed as made up of components and interfaces in this way, then the system's behaviour can be made more intelligible. However, a mind is not a component that can be partitioned off from the world in this way according to Haugeland.[10]

Hutchins, I take it, clearly embraces some form of active externalism. DCog thus incorporates what I will call the EXTENDED claim:

> **EXTENDED**: Cognitive processes are not bound by the skin and skull of an individual but may loop out and include bits of the environment as proper parts.

Note that EXTENDED is different from the first part (i) of COMPUTATION. Active internalists can certainly accept that (some) socio-technical systems are computational systems. Adams and Aizawa, for example, argue that DCog is best seen as a theory of "naturally occurring computation" rather than of cognition, on the ground that processes that exhibit the "mark of the cognitive", all occur the brains of people "as a matter of contingent empirical fact". [2, pp. 46, 59]. Rupert takes a similar stance: socio-technical systems may "act as computational systems, of a sort" but there is no explanatory benefits of treating them as cognitive systems [27, p. 392]. One can of course also accept EXTENDED without accepting the socio-technical systems are computational systems.

[7] As the HCI researchers Victor Kaptelinin and Bonnie Nardi [22, p. 205] have noted, DCog seems to be suited for studying certain highly structured socio-technical systems which some kind of overarching system-level goal can be attributed to. Without such system-level goals it becomes difficult to interpret system activity as a form of problem solving.

[8] I write "typically" since occasionally, Hutchins uses the term 'functional system' instead of 'cognitive system'. His broad conception of computation certainly leaves room for important differences between internal and external computational processes. Hutchins can thus argue that even if what happens inside an individual's head is not a component according to the bandwidth criterion, internal and external processes might be different in such a way that only internal processes ought to be called "cognitive".

[9] The Parity Principle: "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." [8, p. 8]

[10] Note that Haugeland is only using the bandwidth criterion negatively to argue that the mind cannot be partitioned off from the body and the world. He is not using it to partition off some other component (made up of bits of brain, body and world), which could be identified with the mind.

2.4 Socio-technical systems and agency

DCog seems to incorporate yet another claim, which takes it even further away from traditional brainbound cognitive science. Hutchins claims that a social-technical system considered as a whole can have cognitive properties of its own. In discussing the navigation team and its tools as a cognitive system, Hutchins [17] attributes several cognitive capacities to this system, such as perception (p. 182), error-detection (p. 182), self-reflection (see p. 182), remembering (p. 196), and confirmation bias (p. 239). Hollan and Hutchins claim that "[f]rom a distributed cognition perspective, goals may be properties of institutions, but need not necessarily be properties of individuals." [21]

I will call this claim, which is independent of EXTENDED, the AGENCY claim:

> **AGENCY**: A socio-technical system can have a form of agency and be the locus of cognitive capacities such as memory, perception and reasoning.

Many will probably take AGENCY to be highly counterintuitive. Given this, should not AGENCY be read metaphorically, as a claim that it might be *fruitful* to view a socio-technical system as *sort of* an agent? To treat a socio-technical system as an agent, it could be argued, is no more misleading than to treat a subsystem in the brain as an intentional system (one that, say, "interprets" incoming information from other neural subsystems).[11] In this vein, Mark Perry suggests that DCog should be seen as a "representational tool for systems analysis, and not as a true description of activity" and system boundaries should be taken as "artificially defined" [25]. Hutchins is not entirely consistent on this issue, but when takes up the issue of whether mentalistic terms such as 'remembering' are only metaphorically applied to socio-technical systems, he argues that they are not [17, pp. 363–4].

Despite the fact that Hutchins [17] is frequently cited in the extended mind debate, only Rupert [27, 28] and Wilson [32] seem to have picked up on the fact that AGENCY is part of DCog.

3 Outline of the argument

COMPUTATION, which is the core of DCog, suggests a radical reorientation in cognitive science. To include the workings of socio-technical systems among the explananda of cognitive science would amount to a significant widening of the discipline's scope. The second part (ii) of COMPUTATION is therefore in need of some kind of defence. While the boundaries of the unit of analysis need not be restricted to the boundaries of cognition, the relevance of the workings of socio-technical systems for our understanding of cognition needs to be argued for or demonstrated in some way.

There seem to be two routes that proponents of DCog can take to defend COMPUTATION, a direct route or an indirect one. The computational processes of a socio-technical system must either themselves be cognitive (the direct route), or else it must be the case that the unit of analysis needed to make sense of the cognitive processes has to be widened to cover the socio-technical system in which the processes are embedded.

I will argue that EXTENDED cannot help establish COMPUTATION, at least not when EXTENDED is arrived at by appeal to the bandwidth criterion. While EXTENDED can be used to motivate the study of socially *extended* cognition, it cannot, or so I will argue, justify treating whole socio-technical systems as cognitive systems. COMPUTATION therefore needs some other (or further) supporting consideration. I will therefore argue that COMPUTATION depends on AGENCY being true. If AGENCY is accepted, then the claim that cognitive science should study socio-technical systems — the second part (ii) of COMPUTATION — follows naturally.

Should AGENCY be accepted? I will not give an answer to this question, but I will argue that the principles that may lead one to accept EXTENDED cannot be straightforwardly carried over into an argument in support of AGENCY. Towards the end of the paper, I will suggest one way in which AGENCY at least can be made intelligible.

4 From EXTENDED to COMPUTATION

Before considering the direct and the indirect route from EXTENDED to COMPUTATION, I want to briefly consider whether COMPUTATION can be established without the means of EXTENDED or AGENCY.

4.1 Embedded cognition

In an attempt to deflate active externalism, or what he calls *the hypothesis of extended cognition* (HEC), Rupert argues that all the empirical results and observations that active externalists appeal to in order to defend their position can be accounted for equally well (or better) by a *hypothesis of embedded cognition* (HEMC).

> **HEMC**: "[C]ognitive processes depend very heavily, in hitherto unexpected ways, on organismically external props and devices and on the structure of the external environment in which cognition takes place." [27, p. 393]

HEMC indirectly takes the study of how organisms interact with tools and their immediate environment into the purview of cognitive science, although, what cognitive science should ultimately explain is the (internal) cognitive processes. An example will be helpful here. In an ethnographic study of cooperative work in the control room of a London Underground line, the sociologists Christian Heath and Paul Luff [13] note how the two personnel in the control room constantly peripherally monitor each other's activities and design their actions not only to achieve the action's primary goal but also to communicate to each other what they are doing.[12] Such multi-tasking and mutual coordination is ubiquitous in all kinds of settings. Yet, how people manage to do this is hardly something that cognitive science has advanced our understanding of very far.

Such mundane but overlooked patterns of interaction are important phenomena that cognitive science arguably ought to investigate. What cognitive abilities and capacities enable people to smoothly engage in such temporally fine-grained social interaction and monitoring? This example certainly suggests that it may be fruitful for cognitive scientists to pay more attention to what is going on in parts of sociology. However, it seems to me to fall short of making the case that the information flow and "behaviour" of the control room system should be taken as a unit of analysis in cognitive science.

4.2 The direct route

Perhaps the bandwidth criterion (which I take Hutchins to be endorsing) can be used to establish COMPUTATION. According to the

[11] Of course, some think that such explanations are very much misleading [4].

[12] Heath and Luff's study was not informed by DCog, but by ethnomethodology, a theoretical framework in microsociology.

bandwidth criterion, if a socio-technical system is not decomposable into components, which interact through relatively well-defined interfaces, but itself interacts with its environment through such interfaces, then cognitive science ought to, it seems, treat that whole socio-technical system as an explanandum.

In the case of Hutchins' navigation team, this would be plausible if the visual input of the landmarks as presented in the alidade and the auditory output of commands to change are low-bandwidth interaction when compared to the interaction happening inside the system. However, this does not seem to be the case. There are clearly some well-defined low-bandwidth interfaces inside the system. For example, the communication of landmark and bearing information over the telephone circuit between bearing takers and the bearing time-recorder and plotter is clearly a low-bandwidth and well-defined one. In addition, the low-bandwidth interaction of the system as a whole with the wider world of the sea is probably a special feature of this particular socio-technical system (Hutchins does not, I think, claim that *all* socio-technical systems are fruitfully taken as objects of study in cognitive science, but if COMPUTATION turns out to be true only of a very small set of socio-technical systems, then the claim is considerably less interesting).

The bandwidth criterion, it should be made clear, is not the only criterion for determining the boundaries of cognition that has been proposed by active externalists. Andy Clark, for example, rejects the bandwidth criterion as a criterion for determining cognitive system boundaries [7, pp. 156–9]. The existence of genuine interfaces between the brain/body and the world does not, he argues, threaten the claim that cases of genuine cognitive extension are fairly common. What is important is instead that people's cognitive performance often results from "rich temporal integration" of internal and external processes and events [7, sect. 2.6, 4.7]. In such cases, the "fine structure [of internal processes and events] has been selected (by learning and practice) so as to *assume* the easy availability of such and such information" from the external world [7, p. 74]. The emergence of such "subpersonal interweaving" [7, p. 240n11] of internal and external threads is (sometimes?) reflected in personal-level experience, such as when a tool or some other bit of the world become "transparent equipment through which you confront a wider world." [7, p. 74]

Clearly, the proponent of DCog cannot rely on personal-level phenomenology to argue for the second part (ii) of COMPUTATION (unless they are willing to claim that socio-technical systems have experiences). What about the subpersonal-level phenomena of rich temporal integration and interweaving? It certainly seems possible in principle that a whole socio-technical system or practice may emergence in such a way that the all the processes that occur in the system are highly dependent on each other and their organisation. Perhaps Hutchins' navigation team and Heath and Luff's control room are actually examples of such systems. The question is if such a subpersonal (or should it be intersubpersonal?) organisation counts for anything in the absence of personal-level phenomena (superpersonal-level phenomena?).

4.3 The indirect route

The indirect route from EXTENDED to COMPUTATION is analogous with the way in which HEMC leads to the adoption of a larger unit of cognitive analysis (without extending the boundaries of cognition). Assuming that EXTENDED is true, might it not be the case that the extended cognitive processes are deeply dependent on environment of the extended cognitive system. Adapting HEMC somewhat, the proponent of DCog might appeal to the following *hypothesis of embedded extended cognition* (HEMEC):

> **HEMEC**: Extended cognitive processes depend very heavily, in hitherto unexpected ways, on props and devices external to the *extended* cognitive system and on the structure of the *wider* environment in which the extended cognition takes place.

If we assume that the dependency that HEC (EXTENDED), HEMC and HEMEC are concerned with is understood in terms of bandwidth (so that two components are heavily dependent on each other just in case they are coupled in high-bandwidth interaction), then it becomes difficult to argue for HEMEC. If one has already accepted EXTENDED on "bandwidth profile grounds", then all the props and devices that are coupled with an agent in high-bandwidth interaction will already be part of that (extended) agent. HEMEC will therefore not help extend the unit of analysis further. Perhaps there is some other (better) way to unpack dependency without relying on bandwidth profiles, which could justify a further widening of the unit of analysis. As I have mentioned, it is possible to argue for active externalism in other ways than by relying on the bandwidth criterion.

5 From AGENCY to COMPUTATION

To motivate the inclusion of socio-technical systems among the explananda of cognitive science, some notion of group agency seems to have to be made cogent. If some socio-technical systems are agents, then it seems plausible that the computational processes in these systems should be thought of as their cognitive processes. Admittedly, this looks like putting the cart before the horse, since AGENCY is arguably in as much need of justification as COMPUTATION. However, I think looking at the relation between COMPUTATION and AGENCY may throw some light on what would be needed in order to show that they are true.

5.1 Subsystemic representations

Arguments for the existence of group agency, or socio-technical system agency, usually appeal to the explanatory benefits of treating groups or socio-technical systems as agents (see [28]). However, as critics are quick to point out, it seems that the behaviour of groups or socio-technical systems — their "agency" — can be reductively explained by appeal to the behaviour of the people that participate in the system and how they communicate among themselves. For example, one can argue that while a whole navigation team is needed to correctly plot the passage of a ship, the knowledge of the ship's position is found in the head of the plotter, never literally on the navigation chart or diffused in the team and its tools. Similarly, while the organisation of the team must be considered when making sense of the actions of its members, it is redundant to attribute agency to the organisation itself.

Rupert [28] argues, correctly in my view, that to make the case of what he calls "group cognitive systems", it must minimally be shown that the representations used in/by such systems are *mental* representations, not merely cultural/conventional representations that sometimes prompt mental representations in the minds of individual group members. Rupert then argues that according to a number of well-known theories of mental representations, the cultural/conventional representations that are propagated in group cognitive systems fail to count as mental representations.

In cognitive science and the philosophy of mind, one commonly distinguishes between the personal-level of explanation and

the subpersonal-level of explanation. Folk psychological accounts of human conduct, often couched in terms of the beliefs and desires, are examples of person-level explanations. Computational and information-processing models in cognitive psychology, on the other hand, are examples of subpersonal-level explanations. I propose that we make a similar distinction when discussing socio-technical systems. Systemic explanations refer to the "behaviour" of the entire system, in terms of its goal for example (e.g. "navigation into port"), while subsystemic explanations refer to the computational processes that occur in the system.

Rupert presupposes that all representations in a group cognitive system are personal-level representations. Now, this is a plausible presupposition, and as far as I know, it is shared by most philosophers who have defended some group agency thesis. Moreover, when proponents of DCog are out on the field, they are supposed to trace the trajectories and transformations of personal-level representations. However, these representations are ultimately of interest in virtue of their functional roles in the socio-technical system they are trying to understand, in virtue of them being *subsystemic* representations. In DCog, public representations thus have a dual role. They are personal-level representations and — when "functionalised" — they are also subsystemic representations.[13]

Now, I tentatively propose, that one way of cashing out the idea of group (or socio-technical) system agency, is in terms of computations over subsystemic representations that are *not* personal-level representations for any member in the system. To understand the details of how such a system "behaves", a reductive explanation is unlikely to be adequate. If there exist such subsystemic representations inside a system it seems that there might be some explanatory benefit in treating the whole system as a cognitive system, as a kind of agent.

6 Discussion

Much of the previous discussion hangs on the idea that there is a proper domain of explananda for cognitive science. This explananda will consist of the behaviour of various cognitive systems, such as human beings and other animals, and more controversially, the behaviour of robots, software agents, or socio-technical systems. In this paper, when I have referred to explanatory targets or explananda, I have primarily done so by appealing to loose intuitions about what cognition is. I take it to be uncontroversial that cognition is at least primarily an activity of biological organisms, and when we want to extend our notion of cognition to other entities, we have to appeal to similarities to these paradigmatic systems.

The intuition, deep-seated in many, that socio-technical systems simply cannot be agents or cognitive systems may have its roots in the fact that socio-technical systems lack many features of biological organisms. Biological organisms are autopoietic systems, "self-producing" systems, that continuously reproduce their own internal components and boundaries. While there have been attempts to apply such concepts from biology to socio-technical systems, such attempts are I think best seen as metaphorical (see [24] for a brief discussion). The fact that it is so easy to extend notions of mind and cognition from a computational perspective, should perhaps be taken as a sign that the perspective is missing something important.

An alternative way of understanding DCog is to read it as a proposal to revise our very concept of cognition. If this is correct, then the objection that DCog does not fit our intuitions about cognition appears moot. Ronald Giere suggests that such a reading is the most charitable one. He argues against the application of everyday mentalistic notions such as 'believing' and 'remembering' to socio-technical systems but he does not find the notion of 'distributed cognition' objectionable since 'cognition' is a term used primarily by specialists: "We are thus free to develop it as a technical term of cognitive science". [10, p. 318] For Giere, a socio-technical system qualifies as a cognitive system simply by producing or outputting knowledge. The socio-technical system of the navigation team and its tools studied by Hutchins thus make up a cognitive system since it repeatedly produces a fix of the ship's position. On Giere's view, the knowledge of the position is found in the head of the plotter (and possibly one or two other persons), but not on the chart or somehow diffused in the system.

It is possible to read Hutchins [17] as proposing such a revision as well. In a way, he points out that the phenomenon that traditional brainbound cognitive science took as characteristic of cognitive processes, namely the sequential manipulation of symbols, actually manifests itself in various socio-technical systems. So if cognitive science is the science of systems that manipulate symbols or process information, then it should look elsewhere than in the heads of individuals. Such a revision of the concept of 'cognition', Hutchins can argue, allows us step inside the cognitive system and observe symbol manipulation directly [17, pp. 128–9]. I have no objection against such a revision in principle. However, one might argue that it is both arbitrary and redundant [23, 5]. After all, frameworks for studying socio-technical systems and modern organisations are already available within the social sciences.

7 Conclusions

Proponents of DCog, whose works sometimes cited as empirical work that reflect active externalism, seem to be pressed to embrace the idea that some socio-technical systems should be considered to be agents. Appeals to a bandwidth criterion for determining the boundaries of cognitive system do not establish that socio-technical systems should be taken as a unit of analysis in cognitive science. However, many active externalists do rely on the bandwidth criterion to determine the bounds of cognition, but rely on other considerations. It is possible that the bandwidth criterion is not the right one, and that a better criterion will in fact show that (many) socio-technical systems are cognitive systems after all.

Finally, I want to note that these are conclusions about DCog as a framework for studying human cognition. DCog is also widely used in Human Computer Interaction (HCI) and Computer-Supported Co-operative Work (CSCW) research. If the background motivation for our use of DCog is to understand a specific work setting and the (potential) role for information technology in it, then it seems unproblematic to focus on the propagation of representational states in the system. Many information technology systems (especially those deployed in organisational settings) are used to create, transform and propagate various representational states.

ACKNOWLEDGEMENTS

I would like to thank Andy Clark, Julian Kiverstein and Matteo Colombo, and two anonymous reviewers for helpful comments. Thanks also to Microsoft Research for financial support.

[13] Clark [6, pp. 292–3] argues, in the context of active externalism, that speech, writing and other "material symbols" play such a dual role in human cognition.

REFERENCES

[1] Mark S. Ackerman and Christine Halverson, 'Considering an organization's memory', in *Proceedings of ACM 1998 Conference on Computer Supported Cooperative Work*, pp. 39–48. ACM Press, (1998).

[2] Fred Adams and Ken Aizawa, 'The bounds of cognition', *Philosophical psychology*, **14**(1), 43–64, (2001).

[3] Morana Alac and Edwin Hutchins, 'I see what you are saying: Action as cognition in fmri brain mapping practice', *Journal of Cognition and Culture*, **4**(3-4), 629–661, (2004).

[4] Max R. Bennett and Peter M. S. Hacker, *Philosophical Foundations of Neuroscience*, Blackwell Publishing, 2003.

[5] Graham Button, 'Review of edwin hutchins' 'cognition in the wild'', *Computer Supported Cooperative Work*, **6**, 391–395, (1997).

[6] Andy Clark, 'Material symbols', *Philosophical Psychology*, **19**(3), 291–307, (2006).

[7] Andy Clark, *Supersizing The Mind: Embodiment, Action, and Cognitive Extension*, Oxford University Press, 2008.

[8] Andy Clark and David// Chalmers, 'The extended mind', *Analysis*, **58**(1), 7–19, (1998).

[9] Nick V. Flor and Edwin Hutchins, 'Analyzing distributed cognition in software teams: A case study of team programming during perfective software maintenance', in *Empirical Studies of Programmers: Fourth Workshop*, eds., Jurgen Koenemann-Belliveau, Thomas G. Moher, and Scott P. Robertson, chapter 5, 36–64, Ablex Publishing Corporation, (1991).

[10] Ronald N. Giere, 'Distributed cognition without distributed knowing', *Social Epistemology*, **21**(3), 313–320, (July–September 2007).

[11] Christine A. Halverson, 'Activity theory and distributed cognition: or what does cscw need to do with theories?', *Computer Supported Cooperative Work*, **11**(1-2), 243–267, (2002).

[12] John Haugeland, 'Mind embodied and embedded', in *Having Thought: Essays in The Metaphysics of Mind*, chapter 9, 207–237, Harvard University Press, (1998).

[13] Christian Heath and Paul Luff, 'Collaborative activity and technological design: Task coordination in london underground control rooms', in *Proceedings of Second European Conference on Computer-Supported Cooperative Work*, eds., Liam Bannon, Mike Robinson, and Kjeld Schmidt, 65–80, Kluwer, (1991).

[14] James Hollan, Edwin Hutchins, and David Kirsh, 'Distributed cognition: Toward a new foundation for human-computer interaction research', *ACM Transactions on Computer-Human Interaction*, **7**, 174–196, (2000).

[15] Susan L. Hurley, *Consciousness in Action*, Harvard University Press, 1998.

[16] Edwin Hutchins, 'The technology of team navigation', in *Intellectual Teamwork: Social and Technological Foundations of Cooperative Work*, eds., Jolane Galegher, Robert E. Kraut, and Carmen Egido, chapter 8, 191–220, Lawrence Erlbaum Associates, (1990).

[17] Edwin Hutchins, *Cognition in The Wild*, MIT Press, 1995.

[18] Edwin Hutchins, 'How a cockpit remembers its speeds', *Cognitive Science*, **19**(3), 265–288, (July 1995).

[19] Edwin Hutchins, 'Response to reviewers', *Mind, Culture, and Activity*, **3**(1), 64–68, (1996).

[20] Edwin Hutchins, 'The distributed cognition perspective on human interaction', in *Roots of Human Sociality: Culture, Cognition and Interaction*, eds., Nick J. Enfield and Stephen C. Levinson, chapter 14, 375–398, Berg Publishers, (2006).

[21] Victor Kaptelinin, Bonnie Nardi, Susanne Bodker, John Carroll, Jim Hollan, Edwin Hutchins, and Terry Winograd, 'Post-cognitivist hci: second-wave theories', in *CHI '03: CHI '03 extended abstracts on Human factors in computing systems*, pp. 692–693, New York, NY, USA, (2003). ACM.

[22] Victor Kaptelinin and Bonnie A. Nardi, *Acting with Technology: Activity Theory and Interaction Design*, MIT Press, 2006.

[23] Bruno Latour, 'Cogito ergo sumus! or psychology swept inside out by the fresh air of the upper deck… (review of ed hutchins' cognition in the wild)', *Mind, Culture, and Activity*, **3**(1), 54–63, (1996).

[24] John Mingers, 'An introduction to autopoiesis–implications and applications', *Systems Practice*, **2**(2), 159–180, (1989).

[25] Mark Perry, 'The application of individually and socially distributed cognition in workplace studies: two peas in a pod?', in *Proceedings of European Conference on Cognitive Science*, pp. 87–92, (1999).

[26] Yvonne Rogers and Judi Ellis, 'Distributed cognition: an alternative framework for analyzing and explaining collaborative working', *Journal of Information Technology*, **9**, 119–128, (1994).

[27] Robert Rupert, 'Challenges to the hypothesis of extended cognition', *Journal of Philosophy*, **101**(8), 389–428, (2004).

[28] Robert Rupert, 'Minding one's cognitive systems: When does a group of minds constitute a single cognitive unit?', *Episteme*, **1**(3), 177–188, (2005).

[29] Aleksandra Sarcevic, Ivan Marsic, Michael E. Lesk, and Randall S. Burd, 'Transactive memory in trauma resuscitation', in *CSCW '08: Proceedings of the ACM 2008 conference on Computer supported cooperative work*, New York, NY, USA, (2008). ACM Press.

[30] Herbert A. Simon, *The Sciences of the Artificial, 3rd Edition*, MIT Press, 1996.

[31] Robert A. Wilson, *Boundaries of the Mind: The Individual in the Fragile Sciences*, Cambridge University Press, 2004.

[32] Robert A. Wilson, 'Collective memory, group minds, and the extended mind thesis', *Cognitive processing*, **6**(4), 227–236, (December 2005).

[33] Peter Wright, Bob Fields, and Michael Harrison, 'Analysing human-computer interaction as distributed cognition: The resources model', *Human Computer Interaction*, **15**, 1–42, (2000).

Crude, Cheesy, Second-Rate Consciousness

Joanna J. Bryson[2]

Abstract. If we aren't sure what consciousness is, how can we be sure we haven't already built it? In this article I speak from the perspective of someone who routinely builds small-scale machine intelligence. I begin by discussing the difficulty in finding the functional utility for a convincing analog of consciousness when considering the capabilities of modern computational systems. I then move to considering several animal models for consciousness, or at least for behaviours humans report as conscious. I use these to propose a clean and simple definition of consciousness, and use this to suggest which existing artificial intelligent systems we might call conscious. I then contrast my theory with related literature before concluding.

1 INTRODUCTION

> "If the best the roboticists can hope for is the creation of some crude, cheesy, second-rate, artificial consciousness, they still win." — Daniel Dennett (1994), *The Practical Requirements for Making a Conscious Robot*

While leading a group building a humanoid robot in the 1990s, Rodney Brooks complained about the term *robot brain* [1]. You can have a robot hand or arm or eye or even face. But as soon as you say you have a robot brain people say "That's not a brain." The aim of this article is to make you look at some existing artificially-intelligent systems and say "You know, maybe that *is* robot consciousness."

From experience, I know this is hard to do. I was once sitting in a Cambridge, Massachusetts diner with other postdocs after Dennett had just given a seminar. The other postdocs asserted science would solve consciousness, but not in their lifetimes — not in the next hundred years. Their justification for this statement was that we knew nothing about the topic. Even if we accept this statement as fact (which I don't), they conceded that in the previous ten years there were previously many things that we'd known nothing about and had come to understand well. I believe this and more extreme beliefs about consciousness being unknowable are rooted in strong psychological desires for some aspect of human experience or action to be beyond scientific access. In general, a claim that we are "getting now closer" in science often indicates that in fact the claimant does not like the direction science is currently taking them.

While trying to understand why my colleagues were certain we were so far away from a science of consciousness, I challenged them about how a computer could prove itself conscious. Almost anyone who owns a computer can make it type or even say "I am conscious." Dennett [2] implies that our own empathy should be used to judge the achievement. But teddy bears and pet rocks do this with no intelligence at all, while sadly human history is full of people mischaracterising other people as objects.

[1] University of Bath, United Kingdom, email: j.j.bryson@bath.ac.uk
[2] The Konrad Lorenz Institute for Evolution and Cognition Research, Altenberg, Austria

My colleagues the postdocs said that consciousness was a special sort of self-knowledge, being aware of what you are thinking. But computer programs have perfect access to all their internal states. If you set up a program correctly, you can ask it exactly what line of code — what instruction — it is executing at any time, and precisely what values are in its memory. This is in fact the job of program debugging software, such as an Interactive Development Environment (IDE). IDEs are a common type of program which are not generally considered even to be AI, let alone to be conscious [3].

If consciousness is just perfect memory and recall, then video recorders have it. If consciousness also requires access to process as well as memory, then computers have that access. Possibly some people are committed enough to these definitions that they are already convinced computers can be conscious. But in this article I will not focus on phenomenological theories of consciousness. I will look instead at a recent functionalist theory from philosophy, and relate that theory to what is known about the impact of consciousness on expressed behaviour. From this I will propose a new version of the theory that conscious experience correlates perfectly with a particular sort of search for appropriate action selection. Consciousness is a limited-capacity system for learning about potential connections between context and action. We direct it primarily to situations that are uncertain and immediate, which allows us to optimise our use of this resource in building our expertise in our current environments.

2 MULTIPLE DRAFTS AND CONCURRENCY

One well-known functionalist theory of consciousness is Dennett's multiple drafts theory, which starts from the fact that brains have many things going on in them at one time [4, 5]. In Dennett's more recent model, consciousness is a spotlight that shines on no more than one of these things at a time, at least it only shines brightly on one [6]. But why is the brain doing so many things at once? The reason is because if many processors run at the same time, more can get done quickly. In computer science, this is called *concurrency* [7].

Concurrency is a great strategy for problems that can be taken apart into pieces. But the "hard problem" in concurrency comes when you need to combine all or even some of the answers you find back together again. This is called the problem of *coordination*. For an example, think of bees. A colony of bees can explore a large space around their hive to find flowers by having each bee fly in a random direction. They will explore even more space by using simple rules each bee can know, like "don't fly near another bee". But how much would it help the colony if only one bee finds some really good flowers? When the bees communicate by the waggle dance, a lot of bees have to stop what they are doing to be involved, and one bee has to spend a *lot* of time and energy dancing [8]. When you consider not only the cost to the bees currently engaged in the communicative task, but also the complexity of this behaviour and the time it took

to evolve, you realize the dance must represent a substantial adaptive advantage to the bee colony. Some individuals sacrifice time, and the result is that on average highly-related individuals each have a better chance of finding food and bringing it home [9, 10].

How does this relate to consciousness? I suggest that self awareness only seems a significant part of consciousness because there is a significant portion of the self of which we are *not* aware. Put another way, one of the key attributes of consciousness is that it is a "bottleneck" or constraint — a limit that makes some sub-part of an otherwise uniform whole special. In the bee case, that limiting process is the communication to others when a really good source of food has been found by one bee — the recruitment of others to a single location. This same sort of communicating role has also been suggested for consciousness [11–13], but in fact I will propose a markedly different role in the following sections. First though for this section I want to return to discussing consciousness-like elements in extant AI systems.

Some approaches to artificial intelligence also have concurrent processes which normally operate more or less independently. In AI as in other disciplines such as Psychology or EvoDevo, this decomposition of the whole into some specialised subparts is called *modularity* [14, 15]. Just as in Psychology and EvoDevo, the utility of modularity in AI is that more complicated systems can be developed more simply and operate more quickly [16, 17]. The problem of coordination in AI is called *action selection* [14]. This problem emerges whenever multiple modules are contending for a single resource [18]. An example of a "resource" in this sense can be as simple as physical location. I cannot stand and give a talk at a meeting at the same time as I enjoy myself in a café, so if I want to do both I have to find some sequential ordering for my actions. Another such resource is speech — we can only say one word at a time, so words must be sequenced. And, critically for the Dennett [6] description of his attentional spotlight theory, memory. Apparently, episodic memory is a constrained resource, and only some of the things we are thinking about or perceiving will wind up in it.

3 A FUNCTIONALIST HYPOTHESIS OF CONSCIOUSNESS

Dennett [6] has arrived at the conclusion that the only common characteristic of conscious contents is "the historical property of having won a temporally local competition with sufficient decisiveness to linger long enough to enable recollection at some later time". But the question of course is, competition for what? As Dennett points out later and I will return to in later discussion, one element for humans is *public expression*. If your current thoughts made it so far as to become verbalised, they are now a part of the public awareness. In this case the "local competition" is not only internal but also external — with other speakers. The memory is not only your own but also that of any other hearers.

But most theory of mind focuses on individual consciousness. Here it may be a little harder to see why we are only conscious of one thing at a time. Perhaps the phenomenological experience of sequencing in consciousness indicates consciousness is integral to the other sequencing problem, action selection, which I mentioned before. Norman and Shallice [19] propose that consciousness is a set of extra or special resources which are brought to the problem of sequencing behaviour when the brain is either uncertain about the correct sequence (as in a new context or when working on a new task) or when such sequencing is particularly important (as in when performing a delicate operation.) This theory is very similar to my own, with the exception that I will emphasise uncertainty, not heightened control.

Norman and Shallice (and others) have always been somewhat unspecific about what the "special resources" consciousness brings to such difficult situations might be. I am going to make a specific proposal here, although I won't entirely justify it until later in this article.

My proposal is simple — I think consciousness and episodic memory are the parts of a process for adaptable action selection. This process consists of:

1. fixing an aspect of a behaviour context in the brain, and
2. allowing the brain to search for potential actions that might be best suited to this context.

This sort of action selection is exceptional — most aspects of behaviour are predicted directly by their context and do not need such a process of search. However, because human behaviour is unusually plastic, we spend quite a lot of our time doing this sort of thing, even when the next action is not particularly difficult or pressing. Perhaps due to the tools and concepts provided by language and culture, we can even use consciousness to reason about abstract concepts no immediate sensory correlates. Thus we might think about a work we are writing when driving home when the road itself does not demand full attention.

The model I have just described of interacting attention and action I derived from a model developed by researchers in human vision, Wolfe et al. [20]. The main point of their 2000 article is that when performing a new task, one doesn't learn from that performance when one can use vision rather than memory to guide the behaviour. But my hypothesis depends primarily on an incidental model they describe in that work. This model accounts for the difference in the time it takes to find some visual stimuli compared to others.

Studies that measure the time for processing are called *reaction time* (RT) studies. In vision, if you have a field of dots where some are red and one is blue, you will find the blue one very quickly, and your RT will not depend on how many red dots there are. Similarly, if there are a number of **T**s on a screen and one **L**, you will not have trouble finding the one L, and you will find it quickly no matter how many **T**s there are. *However*, if the screen has many **T**s and many **L**s, and **T**s are both red and blue, but only one **L** is blue, it will take you a relatively long time to find the one blue **L**. Further, the more distracting objects there are (red **L**s or blue **T**s), the longer it will take you to find the blue **L**.

Why is this? Vision researchers have long agreed that part of the answer is because finding an object of a particular colour or simple shape are both things problems that your eyes' concurrent systems can handle more or less by themselves. The different cells in your early visual processing can identify whether they have a blue section or a **T** shape easily, and quickly inform whatever decision system needs to know this. But apparently identifying that something is both blue *and* a **T** cannot be done this way. Wolfe and his colleagues proposed a relatively simple explanation for what happens in this case. One just randomly looks at items with one trait and checks if they also have the other trait, until one happens to look at the right one[3]. So for example, you might just look at anything blue in the field (perhaps returning multiple times to some objects) and eventually you will either see that one is also a **T** or give up. Thus the process

[3] There is an older, more complicated theory involving building a "return inhibition map" once a potential target is recognised as inadequate. Wolfe et al point out this extra mechanism is unnecessary so long as the sampling is truly random.

of recognising and visually targeting blueness or **T**ness is not very conscious, but the process of finding a conjunction, saying "is that *both* blue *and* a **T**" apparently must be.

To try to convince you of my definition of consciousness, I will now describe two more experimental psychology examples. Then I will return to the question of conscious machines. Both of my examples concern something Dennett [6] describes as "imponderable" — consciousness in non-human species.

4 ANIMAL MODELS OF CONSCIOUSNESS

4.1 'Declarative' Memory in Rats

My first scientific interest in animal consciousness came when a colleague made passing reference to declarative memory in a rat. Whether or not rats are aware, I was quite certain they didn't declare anything, which is the definition I'd learned for that term. But there is reasonably good evidence rats have explicit episodic memory. We know this from their behaviour, and from its analogies to humans in similar situations. The humans we can ask about their conscious experience.

In this case, the person who was being asked was Henry Gustav Molaison, then known as patient HM. HM had both of his hippocampuses removed to treat his severe epilepsy, and as a result lost the ability to form new episodic memories. When I was a psychology undergraduate in the 1980s, we were taught that he had lost the ability to *consolidate* short-term memories into long-term memories, but this theory proved false. At that time it was believed that when rats had their hippocampuses lesioned (destroyed) they could still consolidate their memory, but they had certain problems with navigation, so apparently hippocampuses were for navigation in rats but memory consolidation in humans. This was also wrong — the real answer is both more parsimonious and more interesting.

What HM can't do is that he can't remember an episode after that episode finishes. So you might teach him one task which he would perform successfully, but then if you distract him by going away or introducing a new task, he could not remember even having met you afterwards, let alone that you had taught him the first task. But although he had his surgery in the 1950s, HM started acquiring semantic knowledge about John F. Kennedy and rock music. Eventually, someone stopped asking HM what he remembered, and instead gave him the same sort of task the lesioned rats were successfully learning. They brought in an apparatus and said "when that light goes on, push that button". When he did so they gave him a penny. After he'd done this for some time, they distracted him by asking him to count his pennies. After this he said he didn't know what the apparatus was for. But when the light went on, he pushed the button, just as a rat would have. When they asked him why he did that, he said "I don't know."

So now that we know that rats and humans were less different than once thought, let us return the question of rat episodic memories. One of the "navigational" tasks the rats had problems with was the radial arm maze — a maze with eight arms coming out from a centre. The trick with this maze is to remember which three arms the scientists put food in, and to go to each of them and not the others because you only have a little time in the maze. Also, you can't learn to go to the three arms in a particular order, because little doors slide up and down randomly, preventing access at irregular times. The rat thus has to remember which of the three arms you've already been down *today* to make sure to go down each of them once. When the rats had no hippocampuses, they could still learn which three arms had the food day after day, just like HM could tell you about the Beatles. But on any particular day, they didn't efficiently go down those three arms once each, like a normal rat would. Rather, they acted like they couldn't remember what they'd just been doing. Just like HM. This is what my colleague had referred to as "declarative memory". The ordinary rats (the ones that still had their hippocampuses) were showing they had it by going down the three arms each once.

For details and full referencing of the above experiments, see Carlson [21]. But the main point here for my argument, is that rats seem to have a special episodic memory, like humans. Also like humans, rats lose that memory if they lose their hippocampuses.

4.2 Absent-Mindedness in Macaques

From the above I hope we can accept that animals as much like us as rats have at least part of what we normally think of as consciousness, and that they use it for remembering things and choosing their actions. Of course, rat awareness is probably quite different from primate awareness. In a controversial set of experiments, Rolls [22] found evidence that while rats occupy their hippocampuses primarily with information about their present location, primates have more representations of the location they are *looking* at. Thus perhaps a rat is *only* self conscious, while a monkey can think about things at other locations.

I will now move on to the third experimental psychology study, on the effect of aging. One of the standard tasks studied in animal cognition is called *transitive inference*. You may remember this from math — if $A > B$ and $B > C$, then $A > C$. Science has shown surprisingly that many animals (even rats and pigeons) find the $A > C$ inference easily — *if* they can learn the two premises. However, it is very, very hard to learn two different premises involving B, one in which it is good and one in which it is bad. Thus animals (and young children) require a great deal of training to memorise the original, adjacent pairs.

The experiment I am about to describe once again depends on reaction time. There are a number of characteristic effects that happen when animals (including humans) learn a sequence of pairs such as: $A > B; B > C; C > D; D > E; E > F$. One characteristic is that the further apart two stimuli are from each other in that chain, the *faster* the animal is at making their choice. This is called the Symbolic Distance Effect (SDE). So due to the SDE, the reaction time for answering $B?E$ is on average shorter than that for answering $B?D$.

As described earlier, reaction times are normally associated with cognition. Historically, researchers have been trying to discover what computation animals might be performing that does transitive inference yet goes faster as a chain gets longer [23, 24]. But the theory of consciousness I presented above provides a different explanation. My theory predicts that the more uncertain animals are about their next action, the longer they hesitate. This allows their brain to search for a better, more certain solution, using a process like I described above for vision.

I came to this theory for two reasons. One is that I have spent some time researching mistakes children and monkeys make in performing transitive inference, and wound up supporting a model of the underlying process that explains everything *except* the SDE. Therefore I [25] — as well as some other people [26] — think the SDE is not dependent on the transitive reasoning. The second reason is even simpler — the SDE can go away and the animals still perform transitive inference correctly. Rapp et al. [27] have shown that elderly rhesus macaques perform transitive inference more quickly than their juniors and just as accurately. However, they have no SDE. All their

transitive decisions are at the same reaction time, which is faster than *any* of the younger monkey's decisions.

If old monkeys can perform transitive inference without an SDE, do then what is it for? Do the older monkeys pay any penalty? Yes: they don't notice if the rewards change on one of their pairs. Because of an error in their experimental design, Rapp and his colleagues started rewarding all their monkeys on the pair $B?D$ at chance, so most of the monkeys (the younger ones) stopped performing $B > D$ and rather went to chance on choosing B or D. But the old monkeys, who hadn't been hesitating, also didn't notice the change in reward and kept choosing B.

This is just one experiment and there's clearly a lot more work to be done. But I put forward as a hypothesis that the older lab monkeys are more likely to go into "auto-pilot" mode on a simple lab task. This could be adaptive for them, since if they'd lived that long in the wild they'd probably already know how to perform most tasks. Further, they might be losing scramble competitions (the way rhesus macaques forage) to younger, more agile monkeys in their troop [28]. Thus learning is probably less important than speed for elderly monkeys. Of course, we can't be sure that they are performing their transitive inference decisions without conscious awareness, because we can't ask them directly about their memory. But hopefully we will find a way to extend this research into human subjects.

5 DO WE HAVE CONSCIOUS MACHINES YET?

Now I return to the question of whether we have already achieved machine consciousness. Maybe not the full rich human pageantry of narrative with qualia, meta-reasoning and everything, but perhaps what Dennett has called "crude, cheesy, second-rate artificial consciousness" [29, p. 137]. What I have proposed above (taken all together) is that calling something "conscious" requires several things:

1. There must be multiple, concurrent candidate processes for conscious attention.
2. There must be some special process applied to a selected one of these processes.
3. This special process must achieve some function, probably concerning sequencing actions. And,
4. as a side effect, the object of this attention will normally be recorded in episodic memory, at least for a while.

Do any machines meet these criteria? I think probably yes. As pathetic as they are compared to humans or our science fiction, I think many of the humanoid robot systems which engage in dialog with human users and attempt to select objects from table tops can probably be thought of as meeting all these criteria in a crude, cheesy sort of way. Such robots are at MIT, Georgia Tech and the University of Birmingham, to name just a few [30–32].

If you think on a larger, Chinese-room sort of scale for a cognitive system, we might also see AI playing a part in other kinds of consciousness. For example, the Internet employs massive concurrency to create a world-wide database of useful information. If someone wants to act on a piece of that information, they employ a search engine to limit their view of all that data to say ten URLs with context on a single web-page. Under the definition of consciousness above, a page enters the consciousness of the system as a whole at the same time it enters the consciousness of the human being who is doing the final selection of the page to be viewed.

Notice that the browser or search-engine on their own would *not* be conscious, because both require the human to do the actual sequencing. However, the human, the browser *and* the chosen search-engine provider (e.g. Google) all retain explicit memory of the selected Internet item and some summary details about its selection, at least for some time. The browser will use this memory to suggest that page to the person again; the search company will use this memory to make it more likely this page is shown to other people who search, and the human will use the information for whatever they originally intended (or possibly something else). Thus in a way a single action selection mechanism is used concurrently by three different cognitive systems. And I think the two forms of consciousness that have AI elements are not too unlike what Dennett [6] refers to as "the publication competence". They are making public conscious information, and this he describes as the final arbiter of what, for a human, is conscious.

6 WHAT THIS THEORY IS NOT

Note that this theory is entirely agnostic about qualia, self representation and so forth. The phenomena described by Lenggenhager et al. [33] for example could well correlate to the sorts of information frequently used by the conscious search process as part of its action selection.

This work is not identical to the currently-popular Global Workspace Theory (GWT) [11, 13]. As I said earlier, while my theory does relate to some coordinated effort between brain systems, the same could be said of any mental process. But I do not believe that *any* process in the brain is global, for simple reasons of combinatorics [34]. I have recently come to believe that processes like those described by Shanahan [13] could well determine the highest-level task- or goal-selection algorithms in autonomous systems, systems that in animals largely correlate to chemical / hormonal regulation systems, [35, 36]. This is an important part of action selection and also one that may be combinatorially accessible. But it is not the same as detailed, dextrous action selection. Much AI experimentation with spreading-activation systems of action selection has shown that these systems do not scale to any sort of complex action selection such as is displayed by mammals [37, 38].

This is not to say I dislike all or even most of the content of the current GWT as described by [39]. My theory covers a far smaller range of the conscious phenomena, but also an aspect which Baars does not concentrate on. The main purpose for consciousness to Baars is to integrate a large variety of information sources. The main purpose of consciousness for me is to allocate an appropriate amount of time to learning about and searching for the next action. These theories may be perfectly compatible. Baars' mechanisms could well be seen as the *how* of consciousness, and the *why is it like that?* Here my theory has focussed on primarily on the *when* and the *what is it for?*

7 CONCLUSION

The goal of this article has been to convince you that there may already be a robot consciousness, at least to the same extent that there are already robot hands and robot legs. Part of the reason we have trouble understanding consciousness is because the term has origins in folk-psychology and as such covers a large range of phenomena, some of which are probably not particularly related [40]. What I have done here is concentrate on two criteria for consciousness Dennett [6] identifies:

1. that it is something that happens to one candidate process among many, and
2. that it creates a lasting impression in something like episodic memory.

From this I have proposed that consciousness is part of a particular process of action selection — one that is triggered by uncertainty and allows for the exploration and association of new actions in a particular context. This is in contrast to the majority of action selection, which is more-or-less reducible to stimulus-response, possibly also with some automated arbitration [41]. From this I have been able to argue that we can find evidence of consciousness not only in animals but also in *existing* AI systems.

None of my arguments are meant to belittle consciousness in any way, although obviously as a functionalist I am happy if they help demystify it. I am not claiming consciousness is emergent, epiphenomenal or being otherwise antirealist. Rather, consciousness is a central process to the part of intelligent behaviour I am most happy to call "cognitive".

Explaining how something works is by no means the same as explaining it away. Similarly, by disassociating consciousness from mystic ideas of soul I do not deny the central role of a concept of self in current human morality, nor the critical importance of moral behaviour to any social species. Even the crude, cheesy, second-rate artificial consciousnesses I have described are not I think belittled by that description — anything but. I think clarifying our concepts on cognition can help us appreciate the progress we have already made in AI as well as improve our approaches. Hopefully as we develop more informed perspectives on intelligence, we will begin building more useful — and more conscious — cognitive systems.

Acknowledgements

This article was originally commissioned for and presented to The Second Vienna Conference on Consciousness (September 2008) as a discussion of Daniel Dennett's contribution to that meeting. Thanks to John Dittami for the invitation; to Wayne Christiansen and Alejandro Rosas Lopez for comments on an earlier draft; and to Dittami and Dennett for their comments at the meeting. This research is supported by a fellowship from The Konrad Lorenz Institute for Evolution and Cognition Research, Altenberg, Austria. Thanks to the University of Bath for providing me the sabbatical leave to pursue it.

REFERENCES

[1] Rodney A. Brooks and Lynn Andrea Stein. Building brains for bodies. *Autonomous Robots*, 1(1):7–25, 1994.

[2] Daniel C. Dennett. *The Intentional Stance*. The MIT Press, Massachusetts, 1987.

[3] Joanna J. Bryson. Consciousness is easy but learning is hard. *The Philosophers' Magazine*, (28):70–72, Autumn 2004.

[4] Daniel C. Dennett. *Consciousness Explained*. Little Brown & Co., Boston, MA, 1991.

[5] Daniel C. Dennett and Marcel Kinsbourne. Time and the observer: The where and when of consciousness in the brain. *Brain and Behavioral Sciences*, 15:183–247, 1992.

[6] Daniel C. Dennett. Can we really close the cartesian theater? Is there a homunculus in our brain? In John Dittami, editor, *Proceedings of the Vienna Conferences on Consciousness*. University of Vienna Press, 2009. in press, available from the Web.

[7] Michael Sipser. *Introduction to the Theory of Computation*. PWS, Thompson, Boston, MA, second edition, 2005.

[8] Karl von Frisch. *The Dance Language and Orientation of Bees*. Harvard University Press, Cambridge, MA, 1967.

[9] Ellouise Leadbeater and Lars Chittka. Social learning in insects — from miniature brains to consensus building. *Current Biology*, 17(16):703–713, 2007.

[10] Ivana Čače and Joanna J. Bryson. Agent based modelling of communication costs: Why information can be free. In C. Lyon, C. L Nehaniv, and A. Cangelosi, editors, *Emergence and Evolution of Linguistic Communication*, pages 305–322. Springer, London, 2007.

[11] Bernard J. Baars. *In the theater of consciousness: The workspace of the mind*. Oxford University Press, USA, 1997.

[12] Peter Carruthers. The cognitive functions of language. *Brain and Behavioral Sciences*, 25(6):657–674, December 2003.

[13] Murray P. Shanahan. Global access, embodiment, and the conscious subject. *Journal of Consciousness Studies*, 12(12):46–66, 2005.

[14] Joanna J. Bryson. Cross-paradigm analysis of autonomous agent architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, 12(2):165–190, 2000.

[15] Joanna J. Bryson. Modular representations of cognitive phenomena in AI, psychology and neuroscience. In Darryl N. Davis, editor, *Visions of Mind: Architectures for Cognition and Affect*, pages 66–89. Idea Group, 2005.

[16] Richard Samuels. The complexity of cognition: Tractability arguments for massive modularity. In P. Carruthers, S. Laurence, and S. Stich, editors, *The Innate Mind: Structure and Contents*, pages 107–121. Oxford University Press, 2005.

[17] Marc W. Kirschner, John C. Gerhart, and John Norton. *The Plausibility of Life*. Yale University Press, New Haven, CT, 2006.

[18] Bruce Mitchell Blumberg. *Old Tricks, New Dogs: Ethology and Interactive Creatures*. PhD thesis, MIT, September 1996. Media Laboratory, Learning and Common Sense Section.

[19] Donald. A. Norman and Tim Shallice. Attention to action: Willed and automatic control of behavior. In R. Davidson, G. Schwartz, and D. Shapiro, editors, *Consciousness and Self Regulation: Advances in Research and Theory*, volume 4, pages 1–18. Plenum, New York, 1986.

[20] Jeremy M. Wolfe, Nicole Klempen, and Kari Dahlen. Postattentive vision. *The Journal of Experimental Psychology: Human Perception and Performance*, 26(2):293–716, 2000.

[21] Niel R. Carlson. *Physiology of Behavior*. Allyn and Bacon, Boston, seventh edition, 2000.

[22] Edmund T. Rolls. Spatial view cells and the representation of place in the primate hippocampus. *Hippocampus*, 9:467–480, 1999.

[23] Peter E. Bryant and Thomas Trabasso. Transitive inferences and memory in young children. *Nature*, 232:456–458, August 13 1971.

[24] Thomas R. Shultz and Abbie Vogel. A connectionist model of the development of transitivity. In *The* 26^{th} *Annual Meeting of the Cognitive Science Society (CogSci 2004)*, pages 1243–1248, Chicago, August 2004. Lawrence Erlbaum Associates.

[25] Joanna J. Bryson and Jonathan C. S. Leong. Primate errors in transitive 'inference': A two-tier learning model. *Animal Cognition*, 10(1):1–15, January 2007.

[26] Brendan O. McGonigle and Margaret Chalmers. Monkeys are rational! *The Quarterly Journal of Experimental Psychology*, 45B(3):189–228, 1992.

[27] Peter R. Rapp, Mary T. Kansky, and Howard Eichenbaum. Learning and memory for hierarchical relationships in the monkey: Effects of aging. *Behavioral Neuroscience*, 110(5):887–897, October 1996.

[28] Lynne A. Isbell. Contest and scramble competition: patterns of female aggression and ranging behavior among primates. *Behavioral Ecology*, 2(2):143–155, 1991.

[29] Daniel C. Dennett. The practical requirements for making a conscious robot. *Philosophical Transactions: Physical Sciences and Engineering*, 349(1689):133–146, October 15 1994.

[30] Deb K. Roy and Alexander P. Pentland. Learning words from sights and sounds: A computational model. *Cognitive Science*,

26(1):113–146, 2002.

[31] Cynthia. Breazeal, Matt Berlin, Andrew Brooks, Jesse Gray, and Andrea L. Thomaz. Using perspective taking to learn from ambiguous demonstrations. *Robotics and Autonomous Systems*, 54(5):385–393, 2006.

[32] Nick Hawes, Aaron Sloman, Jeremy Wyatt, Michael Zillich, Henrik Jacobsson, Geert-Jan Kruijff, Michael Brenner, Gregor Berginc, and Danijel Skočaj. Towards an integrated robot with multiple cognitive functions. In *Proceedings of the Twenty-Second Conference on Artificial Intelligence (AAAI-07)*, pages 1548–1553, 2007.

[33] Bigna Lenggenhager, Tej Tadi, Thomas Metzinger, and Olaf Blanke. Video ergo sum: Manipulating bodily self-consciousness. *Science*, 317(5841):1096–1099, 24 August 2007.

[34] Joanna J. Bryson. Language isn't quite *that* special. *Brain and Behavioral Sciences*, 25(6):679–680, December 2002. commentary on Carruthers,"The Cognitive Functions of Language", same volume.

[35] Joanna J. Bryson and Emmanuel A. R. Tanguy. Simplifying the design of human-like behaviour: Emotions as durative dynamic state for action selection. In J. Vallverdú and D. Casacuberta, editors, *The Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence*. IGI Global, Hershey, PA, 2009. in press.

[36] Philipp Rohlfshagen and Joanna J. Bryson. Improved animal-like maintenance of homeostatic goals via flexible latching. In *Proceedings of the AAAI Fall Symposium on Biologically Inspired Cognitive Architectures*, 2008.

[37] T. Tyrrell. An evaluation of Maes's bottom-up mechanism for behavior selection. *Adaptive Behavior*, 2(4):307–348, 1994.

[38] Eddy J. Davelaar. Sequential retrieval and inhibition of parallel (re)activated representations: A neurocomputational comparison of competitive queuing and resampling models. *Adaptive Behavior*, 15(1):51—71, March 2007.

[39] Bernard J. Baars. Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. In S. Laureys, editor, *The Boundaries of Consciousness: Neurobiology and Neuropathology*, volume 150, chapter 4, pages 45–53. Elsevier, 2005.

[40] Daniel C. Dennett. Are we explaining consciousness yet? *Cognition*, 79:221–237, 2001.

[41] Tony J. Prescott. Forced moves or good tricks in design space? Landmarks in the evolution of neural mechanisms for action selection. *Adaptive Behavior*, 15(1):9–31, March 2007.

Does Embeddedness Tell Against Computationalism? A Tale of Bees and Sea Hares

Matteo Colombo[+§]

Abstract There is a tendency in the cognitive sciences to emphasize that cognition is embedded in a world and intrinsically embodied. In a sense it is uncontroversial that cognition is embedded. However, it is not obvious what the best explanatory framework to understand embeddedness is. Tim Van Gelder and Randy Beer, among others, take it that embeddedness calls for a dynamics perspective and at the same time tells against computationalism: If we want to understand how brains are coupled with their environments, the dynamicist framework is the best, since the traditional computational one seems to be inadequate to capture this kind of phenomena. This paper gives grounds to reject this argument. By focussing on two case-studies, it argues that, on the one hand, the dynamicist framework is not sufficient to understand embeddedness; on the other, the computationalist framework is resilient and is still necessary even to understand a phenomenon involved in embeddedness such as brain-environment coupling. The conclusion is drawn that embeddedness does not provide telling evidence against computationalism.

1 INTRODUCTION

Presumably, no one would question that cognition is, in a sense, situated (or embedded).[1] The environment where an agent is situated plays an important role in its cognitive activity and behaviour. It is also uncontroversial that the relation between an agent and its immediate environment is of one of ongoing interaction. When I say that a cognitive capacity is situated, or embedded, I intend that the processes that underlie that capacity take place and develop when a coupled system emerges from the complex, real-time interplay between brains, environment (and bodies). The adaptive success of a situated agent depends on the kind of causal coupling between its brain, and environment (and body).

Which explanatory framework is best to understand embeddedness is controversial however. Different approaches to cognitive sciences give different force to the claim that cognition is embedded. Tim Van Gelder [17] for example considers embeddedness as an argument in favor of a dynamical approach to cognitive science and *at the same time* a recalcitrant case for the computational approach. "Dynamical cognition – he claims ([17], p. 623) - sits comfortably in a dynamical world." From this claim alone, however, it doesn't follow that the dynamical framework is the best to understand embeddedness. The aim of this paper is to provide some ground for why a dynamical framework may not be sufficient for *understanding* embeddedness. A computational framework still yields necessary insight to understanding how brains and world interact.

Here's how I plan to proceed. Section 2 delimits the playground of my argument. It clarifies the terminology, and makes it explicit the assumption of the argument. Section 3 recalls two case-studies on associative learning. I take it that learning is a paramount example of cognitive ability where embeddedness matters for the following reason. Learning, broadly understood, depends on the interaction between an agent's brain (and body) and its surroundings. By "observing" its interaction with the environment with which its brain is coupled, an agent gains information that enables it to improve its future decisions. Because it increases the "appropriateness" of a behavioural response to a given class of environments, learning contributes to adaptive behaviour. Strictly speaking, then, the target of my argument is whether dynamicists or computationalists give the best explanation of one example of embedded cognitive capacity, namely learning.[2]

Section 4 elaborates on the case-studies by arguing for the necessity of a computationalist framework (and against the sufficiency of the dynamicist one) to understanding embeddedness. Section 5 concludes the paper by summarizing results so far, and drawing the moral that those who think that a dynamicist framework is sufficient for understanding embeddedness may be mistaken.

2 DYNAMICISM, COMPUTATIONALISM AND EMBEDDEDNESS

Because there are various ways in which different theorists conceive of the key terms, it is too easy to be trapped in terminological misunderstandings and lose sight of the

[+] Dpt. of Philosophy, The University of Edinburgh, UK.
§ CRESA (Center for Experimental and Applied Epistemology), Milan, Italy
Email: M.Colombo-2@sms.ed.ac.uk

[1] Embeddedness is often taken to be one of the hypotheses belonging to a family of approaches to understanding cognition known as "Situated cognition movement" [19] . Since in this paper nothing hinges on the distinction between situatedness and embeddedness, I will use 'situatedness' and 'embeddedness' interchangeably.

[2] The assumption that learning is a cognitive capacity that qualifies as embedded is not obvious however. Learning, in fact, covers a diversity of cognitive capacities supported by a variety of mechanisms. It may be that in some cases of learning the role played by the environment is not crucial. And therefore the issue of embeddedness would be something as a red-herring in those cases. Although I shall not put forth a general argument for why, and in what cases, learning qualifies as embedded, I shall motivate why the cases exposed in section 3 can be regarded as cases where embeddedness, or features that characterize embedded cognition (e.g. coupling) are in fact important.

substantial issues in the "dynamicism - computationalism debate". It is therefore necessary to make some terminological clarifications. This section is devoted to put some bits of terminology in place.

Dynamical system theory is a mathematical theory that provides us with analytical tools to study the behaviour of complex systems. Taking a dynamicist perspective on a certain *explanandum* means to adopt specific concepts, metaphors, a vocabulary to frame our understanding of that phenomenon. The core concepts that form the dynamicist framework are the ideas of state space, time set, and evolution operator. These concepts enable us to draw a *geometrical* analysis of the phenomenon we seek to explain. Let a system be a set of interdependent variables that define an *explanandum*, the state of the system at a time is the value of its variables at that time. The state space of the system consists of the possible overall states of the system through time. The state space can be of any topology and dimension according to the number and nature of the variables of the system. The behaviour of the system is understood in terms of the trajectory of the system within the state space through time. The trajectory is governed by the evolution operator which is typically a set of differential equations. Concepts from dynamical systems theory are becoming increasingly important in cognitive science [2]. Dynamical system theory tools are for example among the nuts and bolts of computational cognitive neuroscience [12]. Nevertheless, this does *not* mean that the dynamical *framework* has replaced, or is even near to replace, the "old-fashioned" computational one.

The theory of computation is also a body of mathematics. The adoption of the analytical tools of the theory of computation to study cognition and behaviour of an agent is said 'computationalism'. More precisely, computationalism holds that the cognitive processes and behaviour of an agent is explained by computations. A computation is the manipulation of symbols according to some algorithm. An algorithm is a step-by-step procedure for accomplishing something. A symbol is a carrier of information, it represents something. A related, but conceptually distinct, broader, construal views computation as information-processing. Accordingly, computation is what is involved in manipulating, storing, retrieving, encoding, trafficking information – whatever it might be. If we apply this framework to cognition, cognition and behaviour of an agent are understood in terms of patterns of information transformed, retrieved, stored, processed by a mechanism according to some input-output function. This is what I mean with "computational framework".

A number of disclaimers are in order at this point. First, as mathematical theories the dynamical and the computational theory are roughly equipotent [3]. Therefore, the claim that dynamicism is a better framework than computationalism for understanding embeddedness is not supported by mathematical reasons alone. Second, the characterizations above are "cartoon" characterizations in two senses. On the one hand, they overlook distinctions that motivate different positions within the same approach. For example, Van Gelder's dynamical hypothesis [17] has raised criticism not only from the computationalist party, but also within the dynamicist group. On the other hand, both dynamicism and computationalism face theoretical and methodological "internal" challenges (e.g. [11]; [2], p.114-117). For my purpose however, suffice the cartoon characterization. Third, I assume that *connectionism* represents a refinement of computationalism *and* is in continuity with dynamicism.[3] On the one hand, in connectionism, computations can be seen as distributed across neural networks. On the other, connectionist models can implement equations from dynamical system theory. Finally, and importantly, I take it for granted that cognition is always embedded in an environment. That is, I assume that biological nervous systems are always coupled with an uncertain and changing environment: The brain and the world are causally coupled, and both contribute (with the body) to intelligent behaviour, and cognition. Notice that these are *ontological* claims. Which *explanatory perspective* is the best to understanding embeddedness is a separate issue.

I am not interested in the *nature* of the contribution of the world to cognition. I am interested in assessing whether embeddedness may constitute a really telling argument for assuming a certain *perspective* in the cognitive sciences. A really telling argument for the claim that perspective *X* is better than *Y* (in our case, that dynamicism is better than computationalism) is an argument that supports perspective *X*, and that *at the same time* tells against the rival perspective. If it turns out that computationalism misrepresents embeddedness, whereas the dynamicist framework renders an insightful image of the same phenomenon, then we have an excellent reason to withdraw the computational framework in favor of the dynamicist for understanding embeddedness.

With 'perspective', or 'world-view' - as adopted by Beer [1], I mean a conceptual framework to understanding a phenomenon. A framework comprises a vocabulary, a set of concepts, metaphors, insights that enable us to understand a certain phenomenon. Thus, the computational framework draws on the metaphor of the brain as a computer, and emphasizes the role of functional information-processing structures. Its vocabulary comprises concepts like "manipulation", "processing", representation", "retrieval", "storing", "activation", "input-output function". The dynamicist framework draws on the metaphor of cognition as movement. Its vocabulary comprises concepts like "attractors", "transients", "coupling", "bifurcation", "emergence", "state space", "trajectory".

We should be careful in spelling out what the opposition between dynamicism and computationalism amounts to here. As mentioned above, the opposition is not mathematical, and has not to do with mutual inconsistency. The opposition, instead, is about the different kind of "explanatory priorities", or of "explanatory concerns" suggested by the two frameworks [4], [6], [7]. Taking a computationalist perspective to understanding some system means to focus on the function that the system computes, to try and identify how certain states of the system stands-in, or represents some states of affairs, what class of algorithms transforms these representations, what constraints there are on the information-processing of the system. A dynamical perspective, in contrast, would set a very different class of priorities in one's explanatory agenda. A dynamicist would try to identify the relevant set of variables, and parameters that define the state space of the system. He would be concerned both with the evolution of the system in that state space and with the class of equations that can account for the spatiotemporal

[3] It would be interesting to consider whether connectionist framework is "the best of two worlds" as explanatory framework. Despite its intuitive appeal, the exploration of this possibility goes beyond the scope of this paper.

trajectory of the system. He will be interested in attractors, repellors, phase portraits.

Obviously, *ceteris paribus*, the unfamiliarity of a framework is not a sufficient reason to reject it for another. We have to examine the different kinds of *understanding* provided by the perspectives under scrutiny. To make the point concrete, compare Aristotelian physics with Galilean. In the XVII century, the Aristotelian framework was more familiar than the Galilean one. Its vocabulary comprised "fire, "air", "water", "earth" (four terrestrial elements), and "circular", "up" and "down" (the differential motional natures of the elements). The Galilean framework comprised only one element, "corporeal matter", and different parameters to describe its properties and motions. Facing the same phenomenon, Aristotle talked of a swinging stone striving to reach its natural resting place; instead, Galileo talked of a pendulum, a periodically moving body, whose movement can be understood in terms of frequency, amplitude, radius of the pendulum. The difference is not in precision. The difference is conceptual: By adopting a different vocabulary, the Galilean framework provided better explanatory understanding on the physics of pendula. The ultimate reason why it superseded the Aristotelian framework is empirical: Empirical success always leads the way in the choice of one framework over another.

3 DYNAMICISM AND COMPUTATIONALISM AT WORK

Embeddedness bears on a fundamental cognitive ability: learning. This section centers on two case studies where adaptive behaviour of an agent arises from its ongoing interaction with its environment. Both cases are from the field of computational neuroscience,[4] and both are concerned with learning abilities of simple organisms. The first implicitly assumes that brains are kinds of computers, thereby adopting a computationalist perspective; the second frames its results in dynamicist terms. The two cases in turn.

3.1 HONEYBEES

How may honeybees learn what flowers to visit for getting their next meal? Montague and colleagues [13] tackle this problem by constructing a model of a bee foraging in an uncertain environment. They draw on behavioural observations and neurophysiological data. It seems that bees' foraging behavioural repertoire is based on associations between the occurrence of stimuli (e.g. color of flowers) and outcomes (amount of nectar yielded by the flower). Through trial and error interactions, bees establish proper associations: The more nectar, the more likely bees will return to that flower. Here the key notions to understanding bees' foraging behaviour are *prediction* and *reward*. A bee anticipates what its internal state and the external world will be like by using its current and past experience of reward. *Rewards* are used to improve the quality of predictions. 'Reward' can be defined operationally as the positive value that a system places on the attainment of a certain goal. This kind of learning seems to be supported by a neuron in the bee ganglion, the VUMmx1, which releases octopamine. The activity of this neuron seems in fact to encode a prediction of reward which enables the bee to improve its performances.

Montague and colleagues' model simulates the behaviour of a bee foraging over a virtual field of flowers. The bee is endowed with a visual system that processes inputs from the environment and represents changes in percentages of color. The goal of the bee is to get nectar. To facilitate the attainment of this goal, its computational system guides the bee over the field. The computation is based on current sensory inputs and a prediction of nectar-reward built on remembered rewards associated with different states of the environment, that is, with different colors. The activity of VUMmx1 enables the bee to predict reward by computing prediction-errors with its activity. After the bee has landed on a flower, the neuron combines information about the current reward (the amount of nectar yielded by that flower) with its own prediction (what reward it expected from that flower). Then, it transmits information about how well the actual reward tallies with the predicted one to the rest of the system. When the actual reward is better than expected, VUMmx1 output leads the rest of the system to upgrade the value attached to the state that yielded that amount of reward. In other words, it gives a motive to the bee to remember that the color of that flower predicted an amount of nectar better than expected, and to use this courtesy of the prediction-error system, the bee learns to choose the most adaptive actions by interacting with its environment. Let's turn to another case now.

3.2 SEA HARES

How may sea hares learn to bite edible food and avoid inedible food? Phattanasri and colleagues [15] focus on this food-edibility problem drawing on previous observations of the behaviour of *aplysia*, a genus of sea hares. They model sea hares as agents equipped with a mouth, a smell sensor, and a gut sensor. The goal of this agent is to learn to eat only edible food in a changing environment containing either edible or inedible food. To reach this goal the agent has to learn to associate the right smell to the right type of substance and take an action accordingly by relying on its experiences in that environment.

Phattanasri and colleagues show that a continuous-time recurrent 3-neuron network lacking plastic synapse can evolve to solve this task. This kind of agent is not endowed with a specific learning mechanism; it evolves its learning ability by using the stream of binary smell-sensory inputs from its environment, and the gut sensor serving as a reinforcement signal. Thereby, the evolution of adaptive smell-sensitive actions is function of reinforcement construed as a dynamic property of the [environment-agent-food type] system. Accordingly, they give a topological analysis of the evolution of the system. First, for each of the five possible input patterns (i.e. "no input", "good smell", "bad smell", "positive reinforcement", "negative reinforcement"), the complete phase portrait of the circuit is determined.

The understanding yielded on the learning task under consideration is in fact in terms of phase portraits, that is, of a plot of trajectories in the state space of the 3-neuron circuit. In this way, attractors, basin, and stable equilibrium points are revealed. Then, as the input signal varies over time, the circuit state is observed to move through different phase portraits attracted towards the equilibrium points identified beforehand.

[4] Broadly, computational neuroscience is the use of mathematical modeling, and computer simulations to understand the brain. By itself, this does not entail a commitment to computationalism. However, many computational neuroscientists do make the assumption that brains are kinds of computers (e.g. [5]; [9]).

The dynamic of the system through phase states generates changes in behaviour such that, even though the agent is not endowed with a specific learning mechanism, it can successfully learn to eat only edible food.

4 A PLEA FOR COMPUTATIONALISM

The two cases above serve as bedrock where to build my case for computationalism. My argument is in two parts. First, I examine one reason in favor of the sufficiency of the dynamicist framework, and I rebut it. Then, I motivate why the computationalist framework is still necessary to understand embeddedness.

4.1 ON COUPLING

Let's assume that brains and their surrounding environment are coupled. Coupling is a continuous reciprocal, causal dependence: That a system is coupled with its surroundings means that the system both affects and is affected by what surrounds it [6], [17].

Coupling is usually taken as a reason in support of the arbitrariness of distinguishing brain-centered cognitive systems from the environment where they are embedded [3], [17]. To understand the interactive complexity underlying embeddedness, so runs the argument, we should adopt a dynamicist perspective. In fact, when it comes to *understanding*, relating brain and environment by conceiving of them as a single system is less problematic than relating systems of different kinds. According to this argument, the learning ability of the sea hare is best understood as a dynamics of the system [food-smell sensors-gut sensors] evolving towards an adaptive equilibrium. The general moral is that learning may not be either a behavioural or neural natural kind, but rather, a systemic ability [15]. But does coupling give a strong reason to embrace a dynamicist framework? Or, put differently, does a dynamicist perspective suffice to understand the complex interplay between brains and environment? There are two reasons why it is problematic to affirmatively answer these questions. The first has to do with "componential analysis", the second with "comparative understanding".

First, it may be difficult to understand in dynamical terms the specific, partial contributions of components of the systems to the evolution of the learning ability.[5] The dynamicist may tell us that the contribution of the smell sensors of the sea hare consists in the influence that their values have on the phase portraits in the space state of the system. But this is unsatisfactory: We would like to understand the functional, information-processing role of that component during the evolution of learning. In Montague and colleagues' simulation, we know that learning develops in virtue of an internal supervisor that assesses the ongoing performance of the bee in light of its goals. The bee "is teaching itself about its world" from the feedback of its actions ([16], p.104). Therefore, in accounting for an embedded capacity as learning a dynamical perspective may be insufficient since it would provide little understanding in the information-processing machinery that supports the evolution of the learning of the sea hare.

At this point, the dynamicist may have two objections. He may first point out that it is not obvious how the case of associative learning in the bee is an example of situated cognition where coupling matters. In fact, the bee with its actions doesn't really affect the environment where it is embedded. Why then should we consider this as a case of embedded cognition? In the second place the dynamicist may object that we are simply begging the question: In making this request, we are assuming that the system has to be understood in computational terms. But, according to the dynamicist, it is arbitrary to understanding learning as specifically linked to the information-processing role of one component, rather than to the dynamics of the [brain-body] system.

The response to the first objection goes as follows. The learning of the bee can be considered as an example of situated cognition because the bee does not passively retrieve perceptual information from its world. The representation of the environment that the bee has constantly changes as it makes decisions by drawing upon the values attached to the representations. Thus, certain environmental features (e.g. different patches of color) are really re-constructed depending upon the goal-oriented actions of the bee. By interacting with its world, the bee actively constructs a "value-laden" representation of its environment, which in this sense can be taken to be affected by the activity of the bee. What is the coupled system then? The obvious place to look for causal loops in the bee case is the causal relationship between its value-laden representations of external states of affairs and the decision it makes. The bee's perception, that is, the bee's representation of the external world, affects its decisions which in turn affect perceptions, and so on. Ultimately this kind of complex interplay driven by the prediction-error system leads the bee to display adaptive behaviour.

The second objection can be answered by pointing out to the dynamicist that we have good, *independent* reason to ask about information-processing roles. We regard certain systems as coupled precisely because of that role. A mechanism as gut-sensor is *taken* to be coupled to the sea hare environment because we have a computational pre-understanding of its role. Without having this kind of understanding it would be problematic to identify where to apply the dynamics analysis – whether at the level of brain-body-environment system, or of body-neuromechanical interactions, or neural interactions. For we wouldn't have an *independent* rationale to understanding why we should (de)couple the system in certain ways rather than others. For example, it may be not a good idea to put forth a dynamical analysis of the cognitive ability involved in the conversation you carry on at a crowded pub by focusing on the coupled dynamical system [brain-pub environment]. For in this case it may be more revealing to unfold the computational machinery employed by your brain to pull the right signal out of

[5] Strictly speaking, the sea hare agent modeled by [15] is *non-autonomous*. An autonomous system produces its control signals without benefit of external sensory inputs. Since it receives time-varying inputs, the sea hare agent is not autonomous. However, when it comes to understanding the behaviour of the agent, Phattanasri and colleagues consider it as an autonomous dynamical system by analyzing the sea hare-environment dynamics holding the input fixed to certain values. The coupling here is taken to be a useful *epistemic device* to best understanding the evolution of the agent. One may then wonder whether this study makes a really strong case for dynamicism. Although, perhaps, the sea hare case is not the strongest one for supporting a dynamicist approach to embeddedness, it fits well with my overall argument. In fact, my focus is precisely on *explanatory framework*, on what *epistemic device* is best to understand certain embedded cognitive abilities.

the buzz all around you. For these reasons, it is problematic to take it that coupling is a strong reason to favor a dynamicist approach.

The second problem for a dynamicist framework on embeddedness has to do with the understanding of cognitive analogies across different kinds of agents. Why do agents embedded in different environments (and embodied in very different bodies) seem to display analogous cognitive abilities? Consider the ability of the honeybees to make reward-based predictions and act on its basis. This is an adaptive cognitive strategy that seems to be displayed also by monkeys, rats, and humans. A dynamical framework may be insufficient to understanding this kind of analogy. For, if we treat the agent and its environment as a single complex coupled system, then a cognitive ability has to be understood also in function of the environmental (and bodily) variables of the coupled system. These details are different for bees, rats, humans. Hence, for each kind of agent, we would have different stories about what seems to be the same ability. If this is so, then understanding an apparent analogy would be problematic. In this case, to understand the interplay between brains and environment a unifying, computational framework is best. In fact, a wealth of behavioural and neural data suggests that reward-prediction learning is an ability supported by a particular computation across different kinds of agents. Both in humans, rats, and honeybees, the firing of dopamine (or in bees, a similar chemical called "octopamine") cells seems to encode prediction-error signals governed by a temporal prediction algorithm that accomplishes a specific computational task [14]. A dynamical framework, therefore, may be insufficient to compare abilities across different agents, and thereby grasping analogous cognitive strategies underlying the behaviour of different agents.

4.2 DOING WITHOUT COMPUTATIONS?

The considerations above give some suggestions for why a computational framework is still necessary to understand embeddedness. Let's assume that a computational talk is not only insufficient, but also is unnecessary for an account of embeddedness. We must not talk of the computation of input–output functions in the honeybee case. We must refrain from construing the functions being computed in terms of representations, that is, in terms of the information content carried by the electric signals travelling on a neural network – for instance, information about food smells, or "appropriate" behaviour in a certain kind of environment. If an array of cells computes a prediction-error, we must not call it computation.

It would be unclear what we would gain by framing our understanding of embeddedness purely in dynamical terms. Certainly we would lose grip on the functional role of the components of the system.[6] Dispensing with computational talk may be problematic for a reason of "expressive convenience" as well. For example, in *Science Without Numbers* Hartry Field shows how it is possible to do Newtonian physics without mathematical statements; however, it would be very inconvenient for a scientist, to say the least, if she did without mathematical statements. This possibility is not a sufficient ground to do without numbers.

It might be objected that in all these cases we are unfairly focusing on decoupled brains thereby suggesting a privileged cognitive role for brain-based cognition. This objection, however, misses the target. On the one hand, if we want to understand how brains relate to the environment where they are embedded, we have to understand how the brain component of the brain-body-world system works. And to understand how brains work, the computational framework still seems necessary. On the other hand, computationalism by itself doesn't entail an individualistic "brain-bound" view of cognition. As argued by Wilson [18] and by Clark & Chalmers [8], computational systems that support cognition can extend beyond the skull. The point made above bears on the necessity of a computational framework; a separate issue is whether this framework is better applied narrowly, to understanding the brain alone, rather than widely, to understanding extended cognitive systems of which brains are components. My argument is not meant to have bearing on this issue.

5 CONCLUSION

The Dynamical System Theory is already part of the toolbox of cognitive science. And for good reason since it is an excellent analytical tool to deal with complexity and to have a geometrical analysis of it. However, it is not obvious that dynamics provides the best *conceptual framework* with which to understand cognitive processes. Some (e.g. [1], [4], [15], [17]) suggest that dynamicism is preferable for the study of embedded agents. Embeddedness – the argument runs – provides reason to prefer dynamics over computationalism as conceptual framework. This paper has tried to challenge this argument by examining two cases where the learning of a simple agent interacting with an uncertain environment is understood within different frameworks.

The main claims made and defended so far are five.

1) Even if cognitive systems *are* dynamical systems, it is not obvious that cognitive systems are best *understood* in dynamical terms.
2) Learning is, in a sense, a paradigmatic case of embeddedness.
3) From 2) it doesn't follow that at least certain cases of learning are best understood in dynamical terms.
4) Embeddedness involves coupling. But coupling is not a sufficient reason to prefer dynamics over computationalism.
5) Computationalism is still necessary to understanding (at least certain aspects) of embeddedness.

The conclusion follows that embeddedness may *not* tell at the same time against computationalism and for dynamics. As always in science, empirical results will tell the last word on this issue. If as a conceptual framework computationalism systematically prejudges, thereby biasing, the answers to empirical questions about embedded cognition, then we will have an excellent reason to prefer an alternative framework.

[6] Another way to make this point might be in terms of Dennett's distinction between the physical stance and the design stance. The dynamical framework fails to capture certain properties of systems that become discernible once we offer computational explanations of the same behaviour. This might be akin to Dennett's point that certain explanations become available once we adopt the design stance that are not available from the physical stance [10]. Thanks to Julian Kiverstein for pointing this out to me.

ACKNOWLEDGEMENTS

A sincere thank you to Andy Clark and Julian Kiverstein who provided useful comments on a previous draft of this paper, and to Danielle Brown, who assisted me with the editing of the essay. Obviously, the usual disclaimers about the remaining errors in the paper apply.

REFERENCES

[1] Beer, R.D. (1995). "Computational and dynamical languages for autonomous agents". In R. Port and T. van Gelder (Eds.) (pp. 121-147).

[2] Beer, R.D. (2000). "Dynamical approaches to cognitive science". *Trends Cog. Sci.* 4:91-99.

[3] Beer, R.D. (2008). *The dynamics of brain-body-environment systems: A status report*. In P. Calvo and A. Gomila (Eds.), *Handbook of Cognitive Science: An Embodied Approach* (pp. 99-120). Elsevier.

[4] Beer, R.D. (in press). *Dynamical systems and embedded cognition.* To appear in K. Frankish and W. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence*. Cambridge: Cambridge University Press.

[5] Churchland, P. S., & Sejnowski, T. J. (1992). *The Computational brain*. Cambridge, MA: MIT Press.

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

[7] Clark, A. (1997)."The Dynamical Challenge" *Cognitive Science* 21:4:1997 p 461-481

[8] Clark, A., & Chalmers, D. (1998). "The Extended Mind". *Analysis* 58:10-23.

[9] Dayan, P (1994). "Computational modelling". *Current Opinion in Neurobiology*, 4:212-217.

[10] Dennett, D. (1987). *The Intentional Stance*. The MIT Press.

[11] Floridi, L. (ed.) (2004). *The Blackwell Guide to the Philosophy of Computing and Information*. Oxford - New York: Blackwell.

[12] Izhikevich, E.M. (2007). *Dynamical Systems in Neuroscience.* Cambridge, MA: MIT Press.

[13] Montague, PR, Dayan, P, Person, C, Sejnowski, TJ (1995). "Bee foraging in uncertain environments using predictive Hebbian learning". *Nature* 377:725-728.

[14] Niv, Y., & Montague, R.P. (2009). "Theoretical and empirical studies of learning". In: PW Glimcher et al, (Eds.), *Neuroeconomics: Decision making and the brain*, Chapter 22, 329-249, Elsevier.

[15] Phattanasri, P., Chiel, H. J., & Beer, R.D. (2007). "The dynamics of associative learning in evolved model circuits". *Adaptive Behavior*, 15: 377-396.

[16] Quartz, S.R., & Sejnowski, T.J. (2002). *Liars, Lovers, and Heroes: What the New Brain Science Reveals About How We Become Who We Are*. New York: Harper Collins Publishers Inc.

[17] Van Gelder, T. (1998). The dynamical hypothesis is cognitive science. *Behva. Brain Sci.* 21:615- 628.

[18] Wilson, R.A., (1994). "Wide Computationalism". *Mind* 101: 351-372.

[19] Wilson, R.A., & Clark, A. (2009). "How to Situate Cognition: Letting Nature Take its Course," in P. Robbins and M. Aydede (eds.) *Cambridge Handbook of Situated Cognition*, Cambridge: Cambridge University Press, Ch. 4.

Inference to the Best Explanation: a comparison of approaches

David H. Glass[1]

Abstract. In the form of inference known as inference to the best explanation (IBE) there are various ways to characterize what is meant by the best explanation. This paper considers a number of such characterizations including several based on confirmation measures and several based on coherence measures. The goal is to find a measure which adequately captures what is meant by 'best' and that also yields the true explanation with a high degree of probability. Computer simulations are used to show that the overlap coherence measure achieves this goal, enabling the true explanation to be identified almost as often as an approach which simply selects the most probable explanation.

1 INTRODUCTION

In many scenarios human reasoning seems to involve producing adequate explanations of the phenomena under consideration. In many artificial intelligence applications, however, reasoning and inference can be carried out without any explicit account of explanation. This naturally raises the question as to how explanations can be extracted from such applications. This is crucial if users are to trust the reliability of the inferences made. In probabilistic systems, for example, users often find it difficult to make sense of the reasoning process unless suitable explanations are available. Unfortunately the automatic generation of explanations requires an adequate account of explanation to be given and this is a notoriously difficult problem.

In addition to extracting explanations for the benefit of the user, explanations can play a more fundamental role in reasoning systems. The system could be designed to generate a range of explanations as part of the reasoning process and then go on to select the best one. This form of reasoning is known as abduction or inference to the best explanation (IBE) and has attracted a great deal of interest in the both artificial intelligence (see [13] and [7] for example) and philosophy (see [14] for example). As well as giving an account of explanation, and thus enabling the system to determine whether a proposed explanation should be accepted as a possibility, IBE requires some mechanism for comparing competing explanations. Clearly it would be useful to have a measure for the quality of an explanation and thus to provide an ordering of competing explanations.

A major difficulty for IBE is that there is no generally agreed account of explanation. Considerable effort has been expended by philosophers trying to overcome difficulties with the deductive-nomological and inductive-statistical accounts of explanation as proposed by Hempel [12]. Salmon, for example, gave an account of explanation in terms of statistical relevance (see his account in [19]). However, Salmon along with other philosophers came to realize that statistical relationships alone could not adequately account for explanation: an adequate account would require causality to be taken into account [20]. Statistical relationships may well be important, but this is because they provide evidence for underlying causal relationships.

One of the difficulties with a causal approach is that the concept of causality is just as problematic as explanation. Nevertheless, there has been considerable attention given to causality recently both in philosophy of science and in artificial intelligence. In particular, work on causality within the context of causal models has given rise to accounts of causality which are both practical and philosophically defensible as discussed in [18, 10]. This has opened up the possibility that explanation can be described in terms of such accounts [11].

In light of these points it is helpful to distinguish two components that are required for a full account of explanation:

a) an account of what constitutes an explanation, and
b) a suitable methodology for comparing competing explanations.

This paper builds on earlier work which drew on recent work on probabilistic accounts of coherence in order to meet requirement (b) by providing a measure to rank explanations [9]. In recent years a number of probabilistic accounts of coherence have been proposed and implications for the coherence theory of justification investigated ([3, 16, 17, 2, 1]). Since there has been general agreement that coherence on its own does not result in a high likelihood of truth, the focus has been on the question of whether coherence is truth conducive so that more coherence gives rise to a higher probability of truth. Olsson [17] presents an impossibility theorem to the effect that there is no truth conducive coherence measure. Bovens and Hartmann [1] also present an impossibility result, but argue that its impact on coherentism can be circumvented by adopting a partial ordering of information sets on the basis of coherence, i.e. in some cases one set can be identified as more coherent than another while in other cases no such comparison is possible.

In [9] coherence was considered as a relation between an hypothesis and the evidence for it. The motivation was to find a measure of coherence which matches our intuitive understanding of the concept and to investigate how such a conception might relate to explanation. A connection between the notions of explanation and coherence was established by noting that a condition for a satisfactory account of the relation "... better explanation than ..." turned out to be essentially the same as a plausible condition for the relation "...more coherent than...". After identifying a suitable measure of coherence, several scenarios were presented to illustrate some advantages of this approach over other accounts of 'best explanation'.

This paper expands on this earlier research in two ways. First, in addition to the measures previously used to quantify 'best explana-

[1] School of Computing and Mathematics, University of Ulster, Newtownabbey, Co. Antrim, BT37 0QB, United Kingdom, email: dh.glass@ulster.ac.uk

tion' several other coherence and confirmation measures are considered as possible alternatives. It should be noted that not all of these measures were proposed as measures for ranking explanations, but they do appear to be plausible candidates nevertheless. Second, a computational approach is adopted so that instead of comparing measures on particular scenarios with specifically selected probabilities, they can be compared over numerous scenarios where the probabilities are selected randomly. This gives a picture as to how well the different measures function on average in terms of identifying the actual explanation that is responsible for the evidence.

The structure of the paper is as follows. Section 2 presents a number of possible ways to compare competing explanations. These approaches are then tested using computer simulations in section 3. Section 4 discusses the relevance of this work for the feasibility of IBE as a mode of inference and section 5 presents conclusions.

2 WHAT IS THE BEST EXPLANATION?

In attempting to provide a methodology for comparing competing explanations it is worth noting Hempel's distinction between potential and actual explanations [12]. An actual explanation is one which, as a matter of fact, explains the explanandum in question. A potential explanation is one which, if true, would be an actual explanation. This section considers different approaches for comparing potential explanations. In this context the goal of IBE can be understood as selecting the actual explanation from the potential explanations. Different forms of IBE arise from which of the approaches is used to select the best explanation from the potential explanations.

It is also worth noting another distinction that has been emphasized by Lipton [14] who distinguished between the loveliest explanation and likeliest explanation. To quote Lipton, "We want a model of inductive inference to describe what principles we use to judge one inference more likely than another, so to say that we infer the likeliest explanation is not helpful ([14], p. 60). It seems that the goal for defenders of IBE is to give an account of 'best explanation' in terms of loveliness and show that a feature of such an explanation will be its likeliness, i.e. high posterior probability.

2.1 Approaches based directly on Bayes' theorem

If the goal of IBE is to provide an account of 'best explanation' that will typically have a high posterior probability, then Bayes' theorem provides an obvious starting point. Suppose that there are n hypotheses H_i where $i = 1, \ldots, n$, then the posterior probability of each hypothesis given evidence E is given by,

$$Pr(H_i|E) = \frac{Pr(E|H_i)}{Pr(E)} \times Pr(H_i), \qquad (1)$$

where all the probabilities are assumed to be conditioned on appropriate background evidence k which has been suppressed in the notation.

As pointed out in [9], the most probable explanation (MPE) approach simply takes the best explanation to be the one with the highest posterior probability. This means that hypothesis H_1 is better than H_2 if and only if

$$Pr(H_1|E) > Pr(H_2|E). \qquad (2)$$

Of course, adopting this approach guarantees that the best explanation will be the one that is most probable given the evidence, but it makes IBE trivial since this success has been achieved simply by defining 'best' as 'most probable given the evidence'. In Lipton's terminology the 'loveliest' explanation has simply been defined as the 'likeliest' explanation.

A second approach discussed in [9] is the maximum likelihood (ML) approach. Taking the first term on the RHS of equation (1) hypothesis H_1 is defined to be better than H_2 if and only if

$$Pr(E|H_1) > Pr(E|H_2). \qquad (3)$$

This has certainly some merit to it since good explanations often do make the occurrence of the relevant evidence highly probable. In fact, ideally an hypothesis will deductively entail that the evidence will occur. The problem, however, is that there is no good reason for thinking that an hypothesis with a high likelihood will also have a high posterior probability unless it also has a high prior probability. Thus, despite its merits, we might expect that in many cases IBE, understood as inference to the hypothesis with the maximum likelihood, will not be a good approach for finding true (or highly probable) hypotheses.

A middle way is provided in [4] who define hypothesis H_1 as better than H_2 if and only if

$$Pr(E|H_1) > Pr(E|H_2) \quad \text{and} \quad Pr(H_1) > Pr(H_2). \qquad (4)$$

This approach is referred to in [9] as a conservative Bayesian (CB) approach. A problem with CB is also pointed out in [9] since there are many cases in which the ML and MPE approaches agree as to which of two hypotheses H_1 and H_2 is best and yet CB fails to order them. For example, the ML and MPE approaches will agree in all cases where the priors of the competing explanations are equal and in many cases where the explanation with the greater likelihood has a lower prior, yet in such cases CB does not provide an ordering.

Before going on to look at other approaches, is is worth pausing to ask whether there is a preferred Bayesian account of 'best explanation'. According to Bayesianism, the rational agent updates her degrees of belief according to conditionalization. For example, if she has a prior probability for an hypothesis H_i of $Pr(H_i)$, then conditionalization requires that after taking evidence E into account her probability should be updated via Bayes' theorem as defined in equation (1) so that her posterior probability for H_i is $Pr(H_i|E)$. If she is required to infer one hypothesis, then it seems that this should be the one that is most probable. If so, it might seem that the MPE approach as expressed in (2) is the preferred Bayesian account of 'best explanation', but this is not necessarily the case. The reason for this is that the Bayesian might wish to maintain that she is interested in the most probable hypothesis but that this need not be the one which is the best explanation; it is probability that is important, not explanation. Furthermore, there is no requirement for the Bayesian to infer one of the hypotheses and so even if the MPE approach is adopted this still does not mean that Bayesianism is a form of IBE.

Indeed, van Fraassen [22] has gone further and argued that Bayesianism and IBE are conflicting approaches. Others [15, 14] have responded by arguing that IBE need not involve any departure from Bayesian probabilities and that explanatory considerations may come into play in implementing Bayesian reasoning. A difficulty with this approach is that it seems to require that the 'best explanation' be *defined* as the most probable explanation. But as noted above this makes IBE trivial since in this case the 'best explanation' is guaranteed to be the most probable explanation by definition. The goal of IBE is to give an account of 'best explanation' that is conceptually distinct from 'most probable explanation' and yet show that the best explanation will often be the one that is most probable. The

aim in this paper is to see whether there is an account of 'best explanation' that will achieve this.

2.2 Approaches based on confirmation theory

A confirmation measure of the degree to which a piece of evidence E confirms an hypothesis H, denoted $c(E,H)$, is a measure which satisfies

(i) $c(H,E) > 0 \;\; iff \;\; Pr(H|E) > Pr(H)$
(ii) $c(H,E) = 0 \;\; iff \;\; Pr(H|E) = Pr(H)$
(iii) $c(H,E) < 0 \;\; iff \;\; Pr(H|E) < Pr(H)$

where Pr is a probability function. Another way of putting this is to say that E confirms (disconfirms) H if and only if there is a positive (negative) probabilistic dependence between E and H. It is important to emphasize that confirmation in the sense used here relates to the impact of the evidence on the probability of the hypothesis rather than simply being the posterior probability of the hypothesis given the evidence. This means that the degree to which E confirms H is a measure of how much *more* probable the evidence E makes the hypothesis H.

A large number of confirmation measures have been proposed in the literature (see for example [5]). Here, only three are considered. First, the ratio measure is given by $r(H,E) = log\,[Pr(H|E)/Pr(H)]$, which can be used to rank explanations such that H_1 is defined to be better than H_2 if and only if

$$log\left[\frac{Pr(H_1|E)}{Pr(H_1)}\right] > log\left[\frac{Pr(H_2|E)}{Pr(H_2)}\right]. \quad (5)$$

However, since $Pr(H_i|E)/Pr(H_i) = Pr(E|H_i)/Pr(E)$, it turns out that this results in an identical ordering of explanations as the ML approach considered in the last section. For this reason, the ratio measure will not be considered further here.

An alternative confirmation measure is the difference measure which is given by $d(H,E) = Pr(H|E) - Pr(H)$ and so enables H_1 to be defined as better than H_2 if and only if,

$$Pr(H_1|E) - Pr(H_1) > Pr(H_2|E) - Pr(H_2). \quad (6)$$

The final confirmation measure considered here is the likelihood ratio given by $l(H,E) = log\,[Pr(E|H)/Pr(E|\neg H)]$ which enables H_1 to be defined as better than H_2 if and only if,

$$log\left[\frac{Pr(E|H_1)}{Pr(E|\neg H_1)}\right] > log\left[\frac{Pr(E|H_2)}{Pr(E|\neg H_2)}\right]. \quad (7)$$

It turns out that when there are only two mutually exclusive and exhaustive hypotheses H_1 and $\neg H_1$ all confirmation measures will agree as to which is the best explanation. This is because if E confirms H_1 then it disconfirms H_2 and so the degree of confirmation E provides for H ($\neg H$) will be positive (negative) for all confirmation measures. This does not apply when more than two hypotheses are being considered.

2.3 Approaches based on coherence

In [9] a case was made for using a coherence measure known as the overlap measure proposed in [16, 8] to rank explanations. For an hypothesis H and evidence E the measure in question is given by

$$C_O(H,E) = \frac{Pr(H \wedge E)}{Pr(H \vee E)} \quad (8)$$

whenever $Pr(H \vee E) \neq 0$ and so the suggestion was to define H_1 as better than H_2 if and only if

$$\frac{Pr(H_1 \wedge E)}{Pr(H_1 \vee E)} > \frac{Pr(H_2 \wedge E)}{Pr(H_2 \vee E)}. \quad (9)$$

Another coherence measure proposed in [6] is given by first defining a confirmation measure as

$$F(H,E) = \frac{Pr(E|H) - P(E|\neg H)}{P(E|H) + P(E|\neg H)} \quad (10)$$

and then using this to define a coherence measure, which we shall refer to as the Fitelson measure, as $C_F(H,E) = \{F(H,E) + F(E,H)\}/2$. Using this measure H_1 can be defined to be better than H_2 if and only if

$$C_F(H_1,E) > C_F(H_2,E). \quad (11)$$

It is worth noting that this measure is a confirmation measure as well as a coherence measure.

It turns out that a further coherence measure, the Shogenji measure, proposed in [21], which is defined for H and E as $\frac{Pr(H,E)}{Pr(H)\cdot Pr(E)}$, provides an equivalent ordering to the ML approach discussed earlier. For this reason it will not be considered further.

3 A COMPARISON OF APPROACHES

The goal in this section is to compare the following approaches to ranking explanations:

(MPE) the most probable explanation approach as expressed in (2),
(ML) the maximum likelihood approach as expressed in (3),
(CB) the conservative Bayesian approach as expressed as in (4),
(DIFF) the approach based on the difference confirmation measure as expressed as in (6),
(LR) the approach based on the likelihood ratio confirmation measure as expressed as in (7),
(OCM) the approach based on the overlap coherence measure as expressed as in (9),
(FCM) the approach based on the Fitelson coherence measure as expressed as in (11).

It is not immediately obvious, however, how to compare the different approaches. Since the goal in IBE is to infer the actual explanation then it would seem that MPE would be the most appropriate approach since it will yield the explanation with the highest probability given the evidence. As has already been pointed out, however, this would make IBE trivial and furthermore MPE does not really seem to capture the notion of 'best explanation' adequately. This can be seen from the fact that an explanation can be the most probable one simply because it has a high prior probability and even though it gives a low probability for the evidence, i.e. has a low likelihood.

Another way to address the issue is to look at what kinds of features make an explanation a good one and then see which approach best takes these into account. This is essentially to consider which measure best captures various explanatory virtues. In [9] a case was made that OCM was to be preferred in this respect to MPE, ML and CB and some scenarios were used to motivate this preference. This way of proceeding is somewhat subjective, however, and undoubtedly a case could be made for some of the other approaches on the list since all of them have their merits.

3.1 Methodology

Here the method for comparing the approaches is rather different. Recall that IBE involves first of all consider the explanatory merits of the potential explanations and then inferring the best one as being true or probably true. Hence, the suitability of IBE as an inductive methodology will depend on how often it enables us to identify the actual (or true) explanation. It will, of course, be impossible to do better in this latter respect than MPE, but as we have seen MPE is inadequate as an account of 'best explanation'. The goal is then to see which of the other approaches best approximates MPE, i.e. which of the other approaches yields the actual explanation most often.

In order to do this, computer experiments have been carried out to see how the various approaches perform. The idea is to take a given number of mutually exclusive and exhaustive hypotheses and randomly assign prior probabilities (adding to one) to them. Random values of the likelihoods $Pr(E|H_i)$ are then attributed to the hypotheses. One of the hypotheses is then selected randomly according to the prior probability distribution and designated as the actual hypothesis. Whether E occurs is then decided randomly based on the likelihood ratio for the actual hypothesis. If E occurs the hypothesis which is the best explanation according to each of the approaches above is identified and if it corresponds to the actual hypothesis this is considered a success, otherwise it is a fail. The entire exercise is then repeated to get an average picture of the performance of each approach. It is also repeated for different numbers of hypotheses.

The procedure is summarised in the algorithm below.

initialize number of hypotheses N and number of repetitions R
for $i = 1$ to R **do**
  set count$_E$ and count$_j$ for each approach to zero
  for $k = 1$ to N **do**
    set the prior probability of hypothesis H_k randomly (ensuring they sum to one)
    set the likelihood of hypothesis H_k randomly
  end for
  select one hypothesis based on the prior probability distribution and designate it the actual explanation H_A
  select whether E or $\neg E$ occurs based on the likelihood of H_A
  if E occurs **then**
    increment count$_E$
    for each approach (MPE) to (FCM) denoted j **do**
      select the hypothesis that is the best explanation H_B
      if H_B = H_A **then**
        increment count$_j$
      end if
    end for
  end if
end for
for each approach (MPE) to (FCM) **do**
  print count$_j$ / count$_E$
end for

3.2 Results

Results were obtained for values of N, the number of competing hypotheses, ranging from 2 to 10. In each case 100,000 repetitions were carried out to ensure that the results were accurate. The results are displayed in Figure 1. The accuracy is the number of cases in which a given measure identifies the actual explanation expressed as a percentage of cases in which the evidence E occurs.

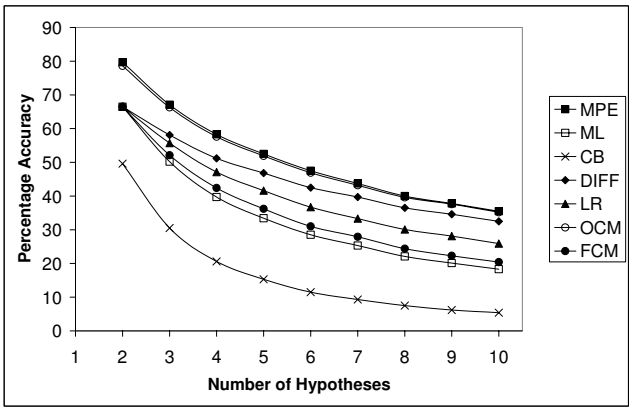

Figure 1. Accuracy plotted as a function of the number of competing hypotheses for each of the different approaches.

It is clear from the results that the percentage accuracy decreases as a function of the number of hypotheses for all of the approaches considered. This, of course, is exactly as expected since as the number of hypotheses increases there are more ways to identify an hypothesis that is not the actual explanation. Also, the MPE approach performs best for all values of the number of hypotheses. This again is as expected since it effectively sets the standard against which the other approaches are to be compared.

Clearly, the CB approach performs worst. In fact, it performs worse than simply selecting an hypothesis at random. For example, when $N = 10$ the accuracy is just 5.4%, whereas 10% could be achieved by random selection. On the other hand, it must be noted that the CB approach is the only one considered here that does not provide a complete ordering of hypotheses and so it also has a much lower false positive rate than other approaches. Nevertheless, it seems clear that the conservative Bayesian approach is too conservative.

The results also illustrate the point, noted in section 2, that all the confirmation measures ML (which is equivalent to the ratio measure), DIFF, LR, and FCM (which is also a coherence measure) yield identical results at $N = 2$, but diverge for $N > 2$. For these higher values of N, DIFF has the best performance, followed by LR, FCM and ML respectively. As N becomes large the results for DIFF get much closer to the MPE results. This seems reasonable since the confirmation measure used in the DIFF approach is the difference measure given by $d(H, E) = Pr(H|E) - Pr(H)$ and so the prior probability $Pr(H)$ becomes less important as the number of hypotheses increases.

Recall that three of the approaches correspond to coherence measures, OCM, FCM and ML (which corresponds to the Shogenji measure, see section 2). Of these, FCM performs slightly better than ML, but OCM outperforms all the other measures. In fact, OCM tracks MPE so closely that it is not possible to distinguish them in the figure. For all values of N the OCM result is within a couple of percent of the MPE result. This means that at $N = 2$, for example, MPE and OCM yield the hypothesis which is the actual explanation almost 80% of the time while the confirmation measures only get the correct result two-thirds of the time and CB half of the time.

Table 1 summarises the performances of the different approaches

averaged over values of N in terms of how well they compare with MPE. The OCM approach does remarkably well, identifying the actual explanation 99% as often as MPE, with DIFF coming in second place with a score of 89%. CB is a long way behind the other approaches with a score of just 30%.

Table 1. The ratio of the accuracy of each approach to the MPE approach averaged over values of N from 2 to 10.

Approach	Average percentage of MPE result
OCM	99
DIFF	89
LR	78
FCM	68
ML	63
CB	30

4 DISCUSSION

The results indicate that the OCM approach using the overlap coherence measure proposed in [16, 8] performs much better than the other measures when compared against the benchmark of the MPE approach. In effect, the OCM approach is almost as good at identifying the actual explanation as the MPE approach. This seems to suggest that OCM does provide a good way of comparing explanations or alternatively a good way of quantifying what is meant by the 'best explanation', but how does this relate to the viability of IBE as an approach to inductive reasoning?

In [9] it was argued that the OCM approach provided a good way of making IBE precise and so IBE can be understood as *inference to the most coherent explanation*, where coherence refers to the coherence between the explanation and the evidence and the coherence measure used is the overlap measure. There it was claimed that it had a number of advantages over MPE, ML and CB and that it provides a good way of linking the goodness of an explanation with its probability of being true without simply defining 'best' as 'most probable'. It was pointed out that any approach which does not define 'best' as 'most probable' will inevitably conflict with MPE in some cases, nevertheless if IBE is to be a viable form of inductive reasoning it should tend to yield explanations that are highly probable. It was claimed that OCM was such an approach, but no experimental evidence was presented.

This paper presents evidence from computer simulations to back up this claim and actually supports a much stronger claim: OCM not only tends to yield highly probable explanations, but it yields the actual explanation almost as frequently as the MPE approach which simply selects the most probable explanation. It is difficult to see how any alternative account of IBE could do better.

This still leaves a question concerning the importance of IBE as an inductive reasoning method distinct to Bayesianism. Two approaches were discussed in [9]. First, perhaps IBE is intended as a descriptive account of how humans actually reason, whereas Bayesianism is the normatively correct way that humans should reason. If so, the rationality of IBE depends on how well it tracks Bayesianism, i.e. how frequently it yields the most probable explanation. As we have seen if IBE is understood as inference to the most coherent explanation, it does remarkably well. Alternatively, perhaps IBE is intended to be a rival to Bayesianism. After all, in some cases the goal of inference is not to find the hypothesis that is most probable given the evidence. As Lipton points out, "... high probability is not the only aim of inference. Scientists also have a preference for theories with great content, even though that is in tension with high probability, since the more one says the more likely it is that what one says is false" ([14], p. 116). Scientists are typically interested in theories which are as precise as possible and testable rather than being vague and compatible with both a piece of evidence and its negation. The account of IBE presented here seems appropriate in this context since the cases where it will diverge from Bayesianism are those when the posterior probability is high because of a high prior and despite a low likelihood.

5 CONCLUSIONS

Various approaches to quantifying the goodness of explanations so that they can be compared and ranked have been considered. These include several simple approaches arising directly from Bayes' theorem, several approaches based on confirmation measures and several approaches based on coherence measures. Results have been presented to show how well each of these approaches performs in terms of identifying the actual explanation. In one sense the MPE approach, which simply identifies 'best explanation' with the explanation that is 'most probable given the evidence', is ideal from an inductive point of view, but it makes IBE trivial and does not provide an adequate account of 'best explanation'. Instead MPE can be seen as the benchmark against which the performance of other approaches should be assessed.

The results show that of the other measures the OCM approach, which uses the overlap coherence measure to identify the 'best explanation', identifies the actual explanation almost as often as the MPE approach. Yet this account seems much more plausible as an account of 'best explanation'. Thus, the research presented here goes some way to vindicating IBE as a form of reasoning provided it is understood as inference to the most coherent explanation where coherence is measured using the overlap measure.

REFERENCES

[1] L. Bovens and S. Hartmann, *Bayesian Epistemology*, Oxford University Press, Oxford, 2003.

[2] L. Bovens and S. Hartmann, 'Solving the riddle of coherence', *Mind*, **112**, 601–633, (2003).

[3] L. Bovens and E. Olsson, 'Coherence, reliability and Bayesian networks', *Mind*, **109**, 685–719, (2000).

[4] U. Chajewska and J. Y. Halpern, 'Defining explanation in probabilistic systems', in *Proceedings of the 13th Conference on Uncertainty in AI*, pp. 62–71, (1997).

[5] B. Fitelson, 'The plurality of Bayesian measures of confirmation and the problem of measure sensitivity', *Philosophy of Science*, **66**, S362–S378, (1999).

[6] B. Fitelson, 'A probabilistic theory of coherence', *Analysis*, **63**, 194–199, (2003).

[7] P. A. Flach and A. C. Kakas, *Abduction and Induction: Essays on their relation and integration*, Kluwer Academic Publishers, 2000.

[8] D. H. Glass, 'Coherence, explanation and Bayesian networks. In Proceedings of the 13th Irish Conference on AI and cognitive science', *LNAI*, **2464**, 177–182, (2002).

[9] D. H. Glass, 'Coherence measures and inference to the best explanation', *Synthese*, **157**, 275–296, (2007).

[10] J. Y. Halpern and J. Pearl, 'Causes and explanations: a structural-model approach-part I: Causes', in *Proceedings of the 17th Conference on Uncertainty in AI*, pp. 194–202, (2001).

[11] J. Y. Halpern and J. Pearl, 'Causes and explanations: a structural-model approach-part II: Explanations', in *Proceedings of the 17th International Joint Conference on AI*, pp. 194–202, (2001).

[12] C. G. Hempel, *Aspects of Scientific Explanation*, Free Press, 1965.

[13] J. R. Josephson and S. G. Josephson, *Abductive Inference: Computation, Philosophy and Technology*, Cambridge University Press, Cambridge, 1994.

[14] P. Lipton, *Inference to the Best Explanation*, Routledge, London, 2nd edn., 2004.
[15] S. Okasha, 'Van Fraassen's critique of inference to the best explanation', *Studies in the History and Philosophy of Modern Science*, **31**, 691–710, (2000).
[16] E. J. Olsson, 'What is the problem of coherence and truth?', *Journal of Philosophy*, **99**, 246–272, (2002).
[17] E. J. Olsson, *Against Coherence*, Oxford University Press, Oxford, 2005.
[18] J. Pearl, *Causality: Models, Reasoning and Inference*, Cambridge University Press, Cambridge, 2000.
[19] W. C. Salmon, *Four Decades of Scientific Explanation*, University of Minnesota Press, 1989.
[20] W. C. Salmon, *Causality and Explanation*, Oxford University Press, Oxford, 1998.
[21] T. Shogenji, 'Is coherence truth-conducive?', *Analysis*, **59**, 338–345, (1999).
[22] B. C. van Fraassen, *Laws and Symmetry*, Clarendon Press, Oxford, 1989.

Noise and bias for free : PERPLEXUS as a material platform for embodied thought-experiments

Olivier Jorand[1], **Andres Perez-Uribe**[2], **Henri Volken**[1], **Andres Upegui**[2], **Yann Thoma**[2], **Eduardo Sanchez**[2], **Francesco Mondada**[3], **Philippe Retornaz**[3]

Abstract. There is a growing interest in attempting to study cognitive and social phenomena under the umbrella of "complex theory". We are indeed immersed in so-called "complex systems", but we are still a long way from a clear understanding of the concepts and principles that underlie the "complexity thinking" [1]. The purpose of this paper to provide a simple (and too short) conceptual framework to understand the basic ideas that allow us to think and speak of complexity in the context of PERPLEXUS as a physical *substratum* for the embodiment of questions related to cognition (individual and/or social) and the material realization of philosophical thought-experiments. To do so, we will notice the controversies concerning the very existence of such a thing as a "theory of complexity". We also will capture some features that can be considered as characterizations (or fingerprints) of "complexity thinking" by contrasting them with a classical Cartesian-Newtonian mode of thinking. Then, we will stress the key role of embodiment as a necessary ingredient to be incorporated in the explanatory efforts of different domains dealing with cognition, development and evolution. We will finally explain how the platform PERPLEXUS can represent such an ideal *locus* for reformatting and tackling conceptual and philosophical questions grounded in aspects of complexity and embodiment.

1 COMPLEXITY : A THEORY ?

Does a "theory of complexity" really exist, or is this expression just a label for a collection of disparate methodologies ? The concept of complexity is often linked today with network science, and researchers wonder if a comprehensive theory with a steady ontological, epistemological and methodological foot is genuinely here. For some, despite its early commercial successes, it will take decades to bring to full fruition what network science provides for an understanding of complexity. For example, Barabasi expressed in 2005 his opinion that: "Despite the necessary multidisciplinary approach to tackle the theory of complexity, scientists remain largely compartmentalized in their separate disciplines. Can they find a common voice ?"[4]. It is a patent fact that "complexity science" (as it is sometimes called) uses in its practical applications both an impressive set of very specialized and technical formalisms (non-linear differential equations, difference equations, networks clustering algorithms, computer simulations to name a few) and less operational, more heuristic guiding principles crystallized in expressions such as "edge of chaos", "emergence" and so on. These ideas have intricate acquaintance with a myriad of others notions such as, higgledy-piggledy: levels of explanation, self-organization, non-linearity, bifurcation, phase transition, fractal, determinist chaos, attractor, dissipative structure, catastrophe, etc., that can make one's head spin. Confronted with the plethora of concepts and terminologies from different disciplines and facing the multiplicity of specific tools and techniques for managing "complex systems" (from now on CS), one could legitimately wonder if it is possible to claim for the existence of a unified theory of complexity. De facto, an "emerging science of complexity" lacks integrated theoretical foundations. In everyday parlance, the expression "CS" is often used to describe an entity that is composed of many interacting parts or components whose structure and behaviour are just plain hard to explain, but even in the systems analysis literature where the adjective 'complex' is ubiquitous, one can find very little to indicate what an author really has in mind when using this terminology. To Casti's eyes for example, the fact is that everyone seems to understand complexity until it is necessary to define it: "In short, we can't really define what we mean by a CS even though we know one when we see it"[2].

Since the question of what constitutes the essence of a CS seems difficult to pin down, Casti thinks that there are actually several facets to the complexity issue depending on the problem, the analyst, the questions being investigated, etc. The pursuit of a viable theory of complexity should take into account theses different facets. We can first discern static complexity which includes inter alia the aspects of hierarchical structure, of connective patterns, of variety of components and of strength of interactions, from dynamical complexity which considers the issues that arise in connection with a system's dynamical motion or behaviour. The different mathematical tools for these aspects are not always naturally related (or even compatible with each other). This is even more the case as soon as we turn to computational complexity which has been approached from different angles too, for example in terms of the size in bits of the shortest program for calculating a binary string (or by extension any digitizable object/phenomenon) in the context of algorithmic theory of information by Kolmogorov-Chaitin [3], or in terms of logical depth by Bennett [4]. These two ways of

[1] Reconfigurable & embedded Digital Systems (REDS), Ecole d'Ingénierie et de Gestion du Canton de Vaud (HEIG-VD), Switzerland.

[2] Institut de Mathématiques Appliquées (IMA), Faculté des Sciences Sociales et Politiques, Université de Lausanne (UNIL), Switzerland.

[3] Laboratoire de Systèmes Robotiques (LSRO), Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland.

[4] Barabasi, A.-L. :"Taming complexity" in Nature Physics, Vol.1, Nov. 2005, p.68-70. Besides, The National Research Council, National Academy Press, Washington DC (2005) reports that "95% of the respondents classify their own work as potentially belonging to network science, yet only 70% claim that such a field exists !".

approaching complexity are associated with two different conceptions of emergence. The former is concerned with synchronic emergence which refers to the identification at a given time of a property present at a given level of a system such as the phenomenon of face recognition; this kind of emergence is associated with a sudden drop of descriptive complexity which allows for a much more concise description of the explanandum than its constituent parts, and which is interpreted in terms of compressibility of information. The latter has to do with diachronic emergence and concerns systems which undergo a process of evolution such as cellular automata where a property (or a pattern, or an object) is considered as emergent if the only way to predict it consist in simulating/running the system by unfolding the scenario of its trajectory from the basic atomic rules deterministically governing its constituent parts. No shortcut possible. This diachronic notion of emergence is obviously illustrated by the increasing degree of complexification of natural entities by the incremental, continuous gradual process of evolution by natural selection.

Many measures of complexity have been proposed for different contexts. However, there is no universal measure that would allow us to establish the degree of complexity of an arbitrary system. Again, each aspect of these distinctions (static, dynamic, computational aspects, …) can be served by different formalisms. The moral is, therefore, that complexity is a multipronged concept that must be approached from several direction keeping in mind the objectives of the analysis. A phenomenon or a system is never universally complex (or complex per se, or complex in an absolute sense). It is complex only in some respects, but not in others. This makes complexity a relative concept, and we now are ready to look in more details some of its constitutive facets, and thus, by first setting the classical Cartesian-Newtonian stage from which complexity thinking detaches itself.

2 TWO PARADIGMS

Based on the above considerations, one can think that a good way of getting a general understanding of complexity thinking is to clarify its principles and concepts by contrasting them with the traditional Cartesian-Newtonian way of thinking. Let's start with a coarse-grained and somewhat caricatural list of contrasting features to then select some of them for further discussion. Firstly, a caveat: we have to keep in mind that this prosthetic list is non-exhaustive, b) that the concepts in each column could be grouped differently, and c) that the columns could be "confronted" differently:

Classical thinking	Complexity thinking
Objectivist theory of knowledge - Strong representationalism - "Naïve" realism	Constructivism – Structural coupling
Top-down analytical approach	Bottom-up synthetic, generative approach
Reductionism - Isolationism	Interactionism – Modularity - Emergentism
Determinism - Predictability	Unpredictability - Non-linearity – Loopyness
Rationalism - Fondationalism	Bounded Rationality
Dualism	Decentralisation, distribution, parallelism, locality
…	Self-organization, adaptation, flexibility robustness, …

Our common-sense understanding of the world alongside with an impressive set of successful scientific models since the advent of modern philosophy and science rely on a classical or Cartesian mode of thinking which is expressed in its most vivid form by Newtonian physics. The ontological and epistemological assumptions of this paradigm that have dominated the scientific view of the world for centuries are -inter alia- a strong representationalist, objectivist, rationalist theory of knowledge which basically establishes a one-one correspondence between the world and our representations of it. Descartes famously codified a top-down notion of analysis consisting in a "divid ut regnat" strategy for conducting reason and seeking truth in sciences via his four principles in his Discours de la Méthode: (1) "never to accept anything as true if I did not have evident knowledge of its truth: that is, carefully to avoid precipitate conclusions and preconceptions"; (2) "to divide each of the difficulties I examined into as many parts as possible"; (3) "to direct my thoughts in an orderly manner, by beginning with the simplest and most easily known objects in order to ascend little by little… to knowledge of the most complex"; and (4) "throughout to make enumerations so complete and reviews so comprehensive, that I could be sure of leaving nothing out". According to this methodological canon, in order to provide a discursive and rational explanation of a phenomenon, one embraces the idea that a whole is a linear combination of its parts, an idea which can be formulated in different idioms such as "superposition" or "compositionality principle".

These principles, when applied to physics, led to the Newtonian materialistic ontology comprising only matter, absolute space and time in which matter moves, and the forces or natural laws that govern these movements; apparently different phenomena are merely different arrangements of separate pieces of matter, of elementary particles ruled by the strict law of cause and effect, leaving no place for intentional, purposeful action unless extended, as Descartes did it, by dualistically postulating an independent category of *res cogitans* completely isolated from *res extensa*. Moreover this reductionist and indefeasibilist way of conceiving a top-down analysis deflates drastically (if not completely) the role played by the interactions between components at the same and different levels of the hierarchical structure of the system under study. The traditional scientific method based on analysis, isolation and the gathering of complete information about a phenomenon is in no position to capture interdependencies between the component parts of an assemblage. Here, what has been called "the laws of nature" deterministically explain both the future trajectory of the system and the path it has taken in the past, implying its predictability and explanation via reversibility. This strict causal determinism finds its standard expression with Laplace in his Essai Philosophique sur les Probabilités: "We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes". Of course, this fondationalist, rationalist view of the universe where an epistemic *Übermench* such as Laplace's

demon can be conceived clashes completely with the "bounded rationality" philosophy of the complexity thinking, where no exhaustive knowledge is at the disposal of faillibilist agents limited in resources and local information, and who have to find good-enough solutions in real-time.

Emancipating itself from this classical paradigm, complexity thinking manifests other "fingerprints" and adopt a set of different and often opposite ontological and epistemological assumptions. The conception of knowledge as passive reflection of the world has not only been questioned in physics by quantum mechanics, relativity theory and by other duality and indeterminacy principles, but also in formal domains such that the foundations of mathematics via formalist or other intuitionist programs and their extensions in verificationist theories of meaning for natural languages. The ontological view of a reality *per se*, as a sort of pure given expecting to be labelled, and the epistemological view of an objective, observer-independent knowledge have been also challenged by numerous developments in cognitive sciences [5] under the general philosophy of constructivism whose motto is best expressed by logician and philosopher Nelson Goodman: "The world is many ways"; cybernetics, biology and embodied cognition have equally shaken the commitment to naïve realism by showing that knowledge is a coevolutionary affair between the knowing subject and the "object", the result of an interactive constructive loop where both pole of the relation co-specify each other. This constitutes a major departure from the reflection-correspondance view of Newtonian epistemology for which the task of science is to refine as much as possible the mapping between the external "reality" and the structures that represent it, be they systems of concepts, images or whatever symbols; in the limit, this mapping should eventually result in a perfect and objective representation of a pre-existing and independent reality, the same for all observers, the understanding of which should be perfect, infallible, reversible and predictable.

As a corollary of this departure from classical reductionism, a recurrent signature of the complexity thinking is, then, the notion of interactionism inducing emergent macroscopic entities -be they properties, objects, processes…a handy general term could be "patterns"- that are non-mystical outcomes resulting from microscopic interactions. The idea is often illustrated by everyday tap water whose properties of being a liquid and non-combustible are emergent properties arising from the interactions of the hydrogen and oxygen "agents" which are both highly flammable gases. We can note at this point that ideas advanced by Conway and Wolfram to avoid a subjective understanding of the concept of emergence consist in showing that there is no shorter path allowing a knowledge of the state of a cellular automata in the future (say after 1000 iterations of the rules) otherwise that the effective applications of their rules; this diachronic characterisation can be formulated more rigorously but, for Wolfram, it suffices to operationalize a subjective notion of emergence rooted in surprise and/or epistemic-cognitive limitations of the observers.

This naturally leads to the fact that the behaviours of CSs are to be understood holistically, i.e. that the global manifested behaviours are the outcome of the multiplicity of its interacting parts whose contributions cannot be detected when taken in isolation. For example, a protein is formed as a chain of amino acids; this one-dimensional sequence of amino acids, strung together like beads on a necklace, specifies how it folds up into a unique three-dimensional configuration that determines its function in the living organism. But it is simply not possible to see how a protein will fold by cutting it at various spots to see how these sub-chains of amino acids fold, and then cementing together somehow the solutions of these individual sub-problems. It must be studied as a single, integrated whole. Relationships between sub-systems turn the whole into a coherent organization with its own identity and autonomy. Actually, the Latin root "complexus", which means something like "entangled, entwined, embraced" analytically contains the idea of components being both distinct and connected, both autonomous and to mutually dependent. Complete dependence would imply order like in a crystal, and complete independence would imply disorder, like in a gas where the state of one molecule gives one no information whatsoever about the state of the other molecules. Etymology indicates us that it is the relations weaving the parts together that turn a system into a complex one, producing emergent properties. Contrary, then, to a complicated TV set whose global functioning can be understood by dissecting it and analyzing its component parts dedicated to one specified function (signal detection, image/sound separation, amplification, etc.), the components of a CS do not implement functions which are totally independent.

That is what Simon [6] has called the property of near decomposability. Near decomposability is a property of a CS, and is two-fold: first, it says that the interactions between sub-systems in a complex system are weaker than the interactions within them (one can think of how interactions between employees in a department are much more frequent than interactions between employees of different departments); second, it says that each sub-system in a decomposed system is almost autonomous, meaning that each is independently functional and useful, but still provides value to the overall system by maintaining a weak connection with it. This kind of near decomposability is reflected in the object-oriented philosophy of programming where loose coupling and encapsulation methods are applied. Thus, CSs can be seen as hierarchical nested structures at different levels of analysis, and what is described as complex at a given level can be understood as a simple component at a higher level. Interactions between such super-systems (that can be seen as agents at the higher level) may recursively produce systems at even higher hierarchical levels, and are a major cause of their unpredictability.

This fingerprint of unpredictability of CSs has been largely popularized under the slogan "butterfly effect" and related concepts such as the sensibility to initial conditions, phase transitions, bifurcations, etc. The common-sense motto "more is different" captures intuitively here the idea that sudden unpredictable new qualitative behaviours can happen once a threshold or a critical mass has been attained. Here, non-linear dynamics and statistical mechanics are the roots of the complexity thinking for dealing with randomness and chaos. The fact that there can be a non-proportional relation between cause and effect can be partly explained by the concept of interaction discussed above: an action induced by a component can cause multiple effects in different parts of the system, and some of these causal chains can close in on themselves. This creation of feedback loops will then either amplify small fluctuations to provoke eventually large global effects by positive feedback, or they can drive and maintain the system in a controlled state

assuring a homeostasis in viable limits by negative feedback, as illustrated by the functioning of a thermostat. Feedback-regulation allows for the emergence of goal-oriented or teleological behaviours that are often describe as being intentional from the outside observer.

3 OTHER FINGERPRINTS

This loopyness is at the heart of a property which is often considered as the hallmark of complexity, namely self-organization [7]: CSs spontaneously organize themselves so as to better cope with various internal and external perturbations, assuring by this their robustness. Fault-tolerance and damage-tolerance make them flexible and guaranty a certain autonomy and adaptability in front of changes of the environment in which they are inserted. No external organizer is required for this organization. Biologists such as Varela and Maturana have discussed at length this kind of autonomy, this "autopoiesis" and this functional interdependence in terms of interactive loop or co-determination between a biological system and its environment/ecological niche. "Order for free" is Kaufmann's slogan for explaining the fact that the property of self-organization can dispense with a notion of an intentional designer. Self-organization can be accelerated by exposing the system to random perturbations, making it visit its state space so that it will reach sooner a state that belongs to an attractor, e.g. the shaking of a pot filled with beans will make them explore a variety of configurations tending to settle into the one that is most stable, i.e. where they are packed most densely near the bottom of the pot, normally reducing their volume. Cyberneticist von Foerster and thermodynamicist Prigogine called this process respectively "order though fluctuations" and "order from noise". This innovative, creative process of self-organization by which the system arranges its components and their interactions into a global structure that tries to maximize its overall fitness without the need of a dedicated controller can be seen as a process of adaptation when we focus on the relations the system has with its environment: whatever the pressures imposed by it, the system will adjust to cope with them. So, evolution can be viewed as the self-organization of an ecosystem into a network of mutually adapted species, and natural self-organization can serve as bio-inspiration as it is the case for genetic and ants algorithms.

This naturally drives us to another directly related fingerprint of complexity, namely the idea of decentralization as vividly illustrated by the decentralized activities of pheromones trails constructions, pigment cells differentiation, fireflies synchronization, applause-bis synchronization, swarm intelligence, etc. There is a feeling of an "invisible hand", one could say, when we witness at a high level of description of a phenomenon the emergence of global patterns which are the outcome of the local properties of the constituent parts without any central organizing force orchestrating the whole process. Decentralized systems are numerous in nature, and one of the distinctive traits of complexity thinking is the abandonment of the "centralized mindset", i.e. the natural tendency for observer to postulate the existence of a cause, a leader, an organizing principle (you name it!) as a decisive causal factor. Complexity thinking reverse completely this view (the famous "argument from design") by adopting the order-for-free or blind-watchmaker view. Indeed, in decentralized, self-organized CSs, no central control is needed for managing, piloting and coordinating the activities of the constituent parts, every one of which has only a partial and limited access to the information computed by the global system. Each element or agent being endowed only with local information, no explicit global description is represented in them. This distribution of information and of competences over the entire system, over all parts or agents constituting it, together with the parallel or asynchronous functioning of the computational resources, are of course the rationale of the robustness, flexibility, fault-tolerance, graceful degradation and other virtues of adaptability of self-organized CSs. Local interactions of the type agents-agents, agents-environment, agents-agents via environment (stigmergy), limited accessibility to information, limited capacity for treatment, etc., can be seen as concepts analytically contained in the idea of distribution, and manifestations of the bounded-rationality principle.

Besides, in complexity thinking computation is substrate-neutral, making the ideas of functionalism and of multiple realizability (no Cartesian nor other forms of dualisms here) parts of its characterization. The Cartesian split between two ontologically incommensurable spheres of being, mind and matter, vanishes: both are particular types of relations. This idea according to which the material substance of a system is irrelevant to the way it performs its function is famously expressed in Bertalanffy's general systems theory [8]: living systems are intrinsically open, i.e. integrate and release information and energy; they therefore depend on an environment so that their effects can never be completely controlled nor predicted. This view is completely different from the traditional Cartesian-Newtonian paradigm in the sense that, ontologically speaking, the building blocks of reality are not to be found primarily in the Newtonian material particles; instead, patterns of organization, i.e. abstract relations, are what are common to different phenomena rather than common material components. Information understood as "a difference that makes a difference", realized on whatever substratum, is what counts. By making abstraction of the concrete substance of components, complexity thinking can establish isomorphisms between systems of different types, noting that the network of relations that defines them are the same at some abstract level, even though the systems at first sight belong to completely different domains. In this context, a super-system imposes a certain coherence on its components, meaning that the behaviour of the parts is to some degree constrained by the whole. This concept of "downward causation" points out for example that the behaviour of an individual is not only controlled by internal neurophysiologic criteria, but also by the emergent regularities of its environment, a point which is of crucial importance in contemporary evolutionary theory of culture.

At that point of our reviewing of the fingerprints of complexity thinking, it is illuminating to remind ourselves of the important historical fact that the concept of self-organization first proposed and developed in the 1940s by the cyberneticists Wiener, Ashby and von Foerster was picked up during the 60s-70s by physicists and chemists studying phase transitions and other phenomena of spontaneous ordering of molecules and particles, and then extended and cross-fertilized during the 80s with the emerging mathematics of non-linear dynamics and chaos. Although, the kind of investigations of CSs practiced by physicists became essentially quantitative and mathematical, a tradition closer in philosophy with the cybernetics origins, developed in parallel

under the heading "complex adaptive systems" with the work of associates of the Santa Fe Institute for the sciences of complexity, among which John Holland, Stuart Kauffman, Robert Axelrod, John Casti, etc. This trend is more qualitative, draws inspiration more from biology and from the social sciences than from physics and chemistry. It is also strongly rooted in computer simulation, and promoted by that the new disciplines of "artificial life" and "social simulation".
This underlines the key role played by spatial structures of interacting agents. These notions of locality, neighbourhood, spatial structures and other dynamical topologies are central ingredients of the dissemination of ideas application of PERPLEXUS, to which we now turn with all these fingerprints in mind.

4 EMBODIMENT AND "ORDER FOR FREE" ON PERPLEXUS

It has been extensively stressed by the robotics community that the synthesis of intelligent behaviour has to transcend the artificial-toy world of pure simulation, and to allow thereby the agent under study to confront its situated body with the for-ever unpredictable contingencies of its environment.
A pretty-fancy program can indeed work without problems in the frictionless and always shining crystal-world of simulation; nevertheless it can, at the same time, not function at all once operating in the real world. The rationale for this phenomenon resides in the ineliminable discrepancies (at all levels of analysis) existing between a simulated model and its real instantiation-implementation : a robot may get stuck against a wall in simulation, whereas it can escape its temporary trap in reality, or vice versa. *De facto*, information-processing tasks that are confined in software abstractions are resolved in ways that are different from the ways employed in real wet life; for example, the real-world structures can be exploited on the fly by cognitive agents without them having a complete-exhaustive representation of it, neither a stock of stored artefactual responses to problems such us collision on the same spot : mobile robots will solve this problem without the need for a conflict resolution scheme that would be, in contrast, needed in a software simulation. As Brooks said in many places : "The world is its own best model".
Although some software tools do integrate today libraries that mimic dynamical properties such as friction, collision, mass, injection of noise, gravity and inertia, etc., these discrepancies between simulation and reality will inevitably cumulate over time : it is a matter of principle that this fact (that could be labelled "the reality gap") is a problem that cannot be resolved without adopting an embodied and situated perspective.
Apart from the increasing recognition in the artificial-intelligence and robotics communities that the nature of the body significantly affects the mind, considerations for supporting an embodied perspective on cognition have had a long story in the biological, ethological, psychological, sociological and philosophical literature : indeed, behaviour is a dynamical process resulting from nonlinear interactions between the agent's control system, its body, and the environment; all these features and the complex patterns of their interactions induce a non linear behavioural trajectory of agent, making its behavior unpredictable although fully determined.
Common theoretical points of the sort (despite differences in terminology), characterizing the "embodied cognition paradigm", will catch the eye of who scans the works of authors from different fields ranging from traditional philosophy (Heidegger, Merleau-Ponty, …), psychology (Vigotsky, Piaget, Thelen, …), ethology (von Üxküll, Gibson, …), biology (Maturana, Varela, …), artificial intelligence (Winograd and Flores Dreyfus, …), robotics (Brooks, Brezeal, Mataric, Beer, Hutchins, Agre and Chapman, Cliff, Harvey, Pfeiffer, Floreano, Mondada, …), neurophilosophy/cognitive sciences (Churchland, Dennett, Clark, van Gelder…), etc.
In this quarters, a generic principled formulation could be the following : intelligent action results in this agent-environment structural coupling which implies a "fuzzification" of clear-cut delineations between mind, body and world as well as perception, cognition and action. Internal world representations that would be complete and explicit representations of the external environment, besides being impossible to obtain and impossible to be used in real time (frame-problem), are not at all necessary for agents to act in a competent manner. To escape the frame-problem, the brain, the body and the world are united in a complex dance of circular causation and extended computational activity (and not considered any more as being clearly separated as in a "Sense-Model-Plan-Act" philosophy typical of a symbolic, a-temporal, static approach which is the typical signature of symbolic AI for example). Emphasis on the physical, environmental, sociological, cultural context reflects the fact that different kinds of minds develop in a given "milieu" and that they use the tools, the representational media, the cultural items, etc., provided by it to support, facilitate, extend, and reorganize mental-cognitive functioning. Here, the relation Subject-World (Agent-Environment, Individual-Ecological Niche, Animal- Umwelt, …) is not simply a one-way (passive) street, but a constructive bi-directional interaction where the agent is a full-blooded constructor of its own behavior and knowledge. The world is here, ready to be "picked-up" and full of connotations of activities decoded according to the needs and possibilities of an embodied agent. This action-centered view of perception is therefore also a plea for embodiment and its indispensable role in reducing the computational burden of cognitive agent.
As a fundamental and defining principle, the embodied cognition paradigm argues that the understanding of the different aspects of cognition rely on explaining them in the context within which the real physical agents operate. That is here that the interdisciplinary project PERPLEXUS [9] enters the scene as an unprecedented opportunity to assess questions that are not ideally addressed in classical computer simulation approaches because of the fact that software abstractions do not do justice to the real anchors of perception-action cycles of embodied cognitive social agents. The project PERPLEXUS aims to develop a scalable hardware platform made of custom reconfigurable devices endowed with bio-inspired capabilities. This platform will enable the simulation of large-scale complex systems and the study of emergent complex behaviours in a virtually unbounded wireless network of computing modules. At the heart of these computing modules, we will use a Ubichip, a custom reconfigurable electronic device capable of implementing bio-inspired mechanisms such as growth, learning, and evolution. These bio-inspired mechanisms will be possible thanks to reconfigurability mechanisms like dynamic routing,

distributed self-reconfiguration, and a simplified connectivity. Such an infrastructure will provide several advantages compared to classical software simulations: speed-up, an inherent real-time interaction with the environment, self-organization capabilities, simulation in the presence of uncertainty, and distributed multi-scale simulations.

Therefore, our agents will have "bodies" and will experience the world, have immediate feedback of their actions on their own sensations so that so that they will be part of a constructive dynamics with their physical environment and their changing social networks. Ubidules-marXbots will operate in dynamic environments using real sensors and effectors and will not be deprived of the possibilities offered by the "world" they live in. Embodiment will assure that they won't get caught in the frame-problem according to which it is by essence impossible to specify a complete model of the world and of the up-dating of its modifications after applications of the operators of its dynamics.

Our embodied societies will be constituted of adaptive agents living in constantly changing environments; more precisely, situated Ubidules riding marXbots able to interpret signals coming from their environment and to communicate thanks to their sensors/effectors equipment, will move around in their environment and disseminate ideas in a non predictable way as function of their perceptual and social biases, and of their constantly changing social networks (dynamical interaction and imitation neighbourhoods). We will profit from the marXbots' perceptual capacities to interact with the real environment and with themselves as a source of the injection of "noises" such as misperceptions, inherent conceptual limitations, interfered transmission, idiosyncratic or socially-influenced preferences to choose these or those (the successful, the common type, etc.) as targets to be imitated.

Although our purpose here is to present the platform as an invitation for researchers to use it as an implementation locus for their own models, we propose in the following paragraphs a possible use of it, just to illustrate its potential exploitation. We think for exemple that questions related to the topics of the dissemination of cultural items will thus find some new opportunities of treatment by the use physical Ubidules-marXbots on the real world platform PERPLEXUS. By making agents embodied and situated, it will be possible to explore aspects of the dissemination of cultural items (and to assess socially-philosophically-oriented questions related to it) that are not ideally addressed in classical computer simulation approaches. From a technical perspective, the embodiment of cultural dissemination mechanisms in a group of mobile robots implies a set of challenging requirement for the mobile robotic infrastructure itself: *1) network size:* we need a sufficiently high network size to achieve emergence and run experiments that are representative for social exchanges. Therefore we need to ensure that experiments will involve at least 20 robots. If this number is not exceptional anymore in the field of collective robotics, it appears to be an interesting challenge when combined with the others requirements of our application in term of flexibility, monitoring and embedded features; *2) complex interface with the environment:* the robots need a sufficient number of sensors and actuators to perform basic tasks combined with some social communication and exchange of ideas. The communication of ideas implies extended possibilities of expression and of perception. The marXbot's design is the result of a long experience in collective robotics, where this topic has been already addressed at a lower scale and exploited in the European projects “swarm-bots” and “ECAgents”; *3) flexibility*: the marXbot's design is based on a modular structure allowing a very efficient adaptation of the functionalities of the robot and will provide the necessary flexibility; *4) experimentation tools*: research in collective robotics is extremely demanding in term of infrastructure to efficiently run experiments, monitor and document them. Controlling the operational condition of 20 robots, monitor their activity, movements and provide the pertinent information to the researcher is a heavy task demanding a specific infrastructure. The marXbot's design takes in account this aspect providing each robot with an onboard LINUX and wireless access. This allows an excellent accessibility to the machines both from the development and the experimentation perspective. The robots will be exploited in an arena equipped with a tracking system allowing an optimal monitoring of the displacements. Because of the compact size of the marXbot, the arena will have a reasonable size and allow a wide range of experiments; *4) duration of experiments:* the systematic exploration of complex emergent phenomena will require experimentation on long periods of time. Energy is a well-known limitation in mobile robotic systems and often sets strong limitations in term of duration of experiments. The marXbot's design includes an energy management system allowing swapping battery during operation. This feature will be exploited to provide several days of autonomy to our mobile robotic system.

With this set of technical features we will be able to embed into a robotic system a set of social interaction experiments exploring emergence of culture in an innovative and efficient way. But we insist on the fact that, from a more general point of view, in additional to an ideal setup for evaluation of culture dissemination (which is just one exploitation among many possible), this setup will be an optimal tool to explore ubiquitous computation in a dynamic network of mobile systems.

5 CONCLUSIONS

We have explained that for some mathematicians and thinkers the notion of complexity, as ubiquitous as is, is nevertheless a multipronged concept and that a system is not complex per se, but deserves the predicate "is complex" only in a relative sense. A full-blooded, unified formalized a theory of complexity does not yet exists and still awaits its Pascal and Fermat. This fact implies that the elaboration of a decent theory of complexity must begin by identifying and analyzing the key components of the kind that we have evoked in this paper. Although humility must win over hubris talking about a comprehensive understanding of the notion of complexity, reasons for optimism are fuelled by an interdisciplinary pursuit towards characterizing these key facets of complexity, formulating organizing principles, making distinctions and clarifying their ontological and epistemological foundations in order to augment our awareness of its multidimensional fingerprints [10].

We have then discussed some of these fingerprints of complexity thinking by contrasting it with a classical Cartesian-Newtonian mode of thinking. This allowed us to underline the transition from a mechanist and deterministic ontological and epistemological view to a more global and modest approach. This modesty hides a new ambition: crating an artificial world, with virtual and embodied agents/societies which manifest

behaviours analogous with the ones we observe in the real world [11]. However, computer simulations are intrinsically limited for capturing what the real word has to say (so to speak) concerning, cognitive downloading and other co-evolving agent-environment phenomena. PERPLEXUS represents an opportunity to gain insights into dynamic processes that standard mathematical techniques would not reveal and that computer simulations would not capture. As a material computational platform, it does overcome these kinds of limitations and can serve as a physical *substratum* for the embodiment of questions related to cognition (individual and/or social) and the material realization of ***philosophical thought-experiments.***

In this sense, PERPLEXUS will represent a unprecedented aid for intuition, imagination, testing as well as a major adjuvant for explorations of ideas concerning multi-secular conceptual and philosophical questions. It is our hope that the family of models developed so far on the pervasive computing infrastructure PERPLEXUS (and whose generality allows for extensions) can humbly serve as ***tools for thinking*** aspects of the deep and important topics of the "embodied cognition" paradigm.

Acknowledgments

PERPLEXUS is funded by the Future and Emerging Technologies programme ISTSTREP of the European Community, under grant IST-034632 (PERPLEXUS). The information provided is the sole responsibility of the authors and does not reflect the Community's opinion. The Community is not responsible for any use that might be made of data appearing in this publication.

REFERENCES

[1] Holland, J.H.: *Hidden Order. How Adaptation Builds Complexity*, Addison-Wesley, New York, 1996.

[2] Casti, J.: *Connectivity, Complexity, and Catastrophe in Large-Scale Systems.* New York: John Wiley and Sons, Inc., 1979.

[3] Chaitin, G.J.: "Information-Theoretic Computational Complexity", *IEEE Transactions on Information Theory*, IT-20, 1974, pp. 10-15.

[4] Bennett, C.H.: "Logical Depth and Physical Complexity", The Universal Turing Machine; a Half-*Century Survey*, Rolf Herken (ed.), OUP, 1988, pp. 227-257.

[5] Jorand, O.: *Pour une Evaluation des Approches Connexionnistes de la Construction des Concepts.* Thèse d'habilitation, Université de Fribourg, Faculté des Lettres, 2006.

[6] Simon, H.: *The Sciences of the Artificial*. MIT Press, Cambridge, Mass., 1st edition 1969.

[7] Ashby, W. R.: "Principles of the self-organizing System" in von Foerster, H. and Zopf, G.W. (Eds*.), Principles of Self-Oganization*. Pergamon Press, 1962, 255-278.

[8] von Bertalanffy, K.-L.: *General System theory: Foundations, Development, Applications*. New York: George Braziller 1968, revised edition 1976.

[9] E. Sanchez, A. Perez-Uribe, A. Upegui, Y. Thoma, J. Moreno, A. Villa, H. Volken, A. Napieralski, G. Sassatelli, and E. Lavarec, "PERPLEXUS: Pervasive computing framework for modeling complex virtually-unbounded systems," in AHS 2007 - *Proceedings of the 2nd NASA/ESA Conference on Adaptive Hardware and Systems*, T. Arslan et al, Ed. Los Alamitos, CA, USA: IEEE Computer Society, aug 2007, pp. 600–605.

[10] Heylighen F./ Cilliers, P./ Gershenson, C.: "Complexity and Philosophy" in Bogg, J. and Geyer, R. (eds): *Complexity, Science and Society*, Radcliffe Publishing, Oxford, 2007.

[11] Volken, H.: "La pensée décentralisée, une innovation majeure dans la modélisation mathématique" in *L'invention dans les sciences humaines*, Bridel, P. (ed.), Labor et Fides, 2004.

Towards a computational model of embodied mathematical language

A. Pease[1], P. Crook[1], A. Smaill[1], S. Colton[2] and M. Guhe[1]
[1]School of Informatics, University of Edinburgh
[2]Department of Computing, Imperial College London
A.Pease@ed.ac.uk

Abstract

We outline two theories of mathematical language acquisition and development, and discuss how a computational model of these theories may help to bridge the gap between automated theory formation and situated embodied agents. Finally, we briefly describe a simple theoretical case study of how such a model could work in the arithmetic domain.

Introduction

It is surprising that little work has been carried out into the way in which humans develop mathematical language, both on an individual and social level. A better understanding of the processes by which we learn to represent, store, communicate, use and develop mathematical ideas would have great educational potential as well as implications for other language acquisition, philosophy and psychology of mathematics, and robotics. The deficiency of work in this area led cognitive scientists Lakoff and Núñez to lament in 2001 that (prior to their work) "there was still no discipline of mathematical idea analysis" (Lakoff and Núñez, 2001, p.*XI*). A philosophical counterpart to Lakoff and Núñez's work is Lakatos's work in the philosophy of mathematics (Lakatos, 1976). Both theories reject the "romantic" or "deductivist" style in which mathematics is presented as an ever-increasing set of universal, absolute, certain truths which exist independently of humans, arguing instead that mathematics uses non-absolute, defeasible reasoning.

Lakoff and Núñez's theory of embodied mathematics

Lakoff and Núñez present the thesis that the human embodied mind brings mathematics into being (Lakoff and Núñez, 2001). That is, human mathematics is grounded in bodily experience of a physical world, and mathematical entities inherit properties which objects in the world have, such as being stable over time. They review studies which suggest that babies are able to distinguish one (small) number from another, to know the size of a small collection of objects (although not necessarily link size to order, so "3" is seen as different to, but not necessarily as bigger than, "4"), and to perform very simple arithmetic (see also (Butterworth, 1999)). For the sake of their argument, these abilities are called innate arithmetic. In order to form more complex mathematical ideas, we need to be able to form two types of conceptual metaphor between innate arithmetic and the more complex arithmetic of natural numbers. Firstly, we need to be able to make *grounding metaphors*. These allow us to project from everyday experiences onto abstract concepts. For instance, we make the metaphor between putting physical objects into groups, and the abstract concept of addition. Lakoff and Núñez identify four grounding metaphors for arithmetic: forming collections, putting objects together, using measuring sticks, and moving through space. The second type of metaphor that we need to be able to make is a *linking metaphor*. This consists of blending different metaphors and yields sophisticated ideas, such as mapping points on a line to numbers, algebraic equations to geometrical figures, or numbers to sets. Lakoff and Núñez argue that much of the abstraction of higher mathematics is the consequence of this type of systematic layering of metaphor upon metaphor and they show where mathematical concepts and laws come from, in terms of these metaphors.

The importance of the environment and our interaction with it in the development of mathematical ideas and capabilities is supported by work in mathematics education and psychology. For instance, Dienes developed a theory of embodied mathematical knowledge and situated cognition, claimed that the environment is "of outstanding importance" in learning mathematics (Dienes, 1973), and, in his theory of the acquisition of mathematics (discussed in (Taylor, 1976)), argued that interaction with the environment is a fundamental aspect of three of the six stages. Piaget gave experience in the environment and action central roles in the developmental process (Piaget, 2001). Choat provides another example in his argument that "all mathematical knowledge originates from contact with objects which constitute the environment" (Choat, 1980, p. 38).

Lakatos's theory of social mathematics

Lakatos charts the evolution of meaning of mathematical terms via dialectic. His influences include Hegel's dialectic, in which the *thesis* corresponds to a naïve mathematical conjecture and proof; the *antithesis* to a mathematical counterexample; and the *synthesis* to a refined theorem and proof (described in these terms in (Lakatos, 1976, pp.144-145)). Another influence is Plato, and some of the reasoning which Lakatos describes can be compared to that in Plato's *Republic*, in which arguments are not deductive: the meaning of terms in the arguments changes over time, and therefore a term in a premise of an argument may not mean the same as the same term in the argument's conclusion. For instance, Simonides proposes that "it is right to give back what is owed". This initial statement is questioned by Socrates with the counterexample of someone borrowing weapons from a friend who subsequently goes insane, in which case it would not be right to return the weapons. The discussion in *The Republic* then turns to what it means to give back what is owed, with Polemarchus suggesting that people owe their friends good deeds, and their enemies bad ones. The dialogue later turns to what the concept of *doing right* means, and leads into Plato's treatment of justice. Another example is the change of meaning of the mathematical term "set" which evolved, in response to Russell's paradox and other problems, from Cantor's "collection of objects" to Zermelo-Fraenkel's definition: "given the set S, and any meaningful property P, it is possible to form the set of all members of S which satisfy P". Lakatos calls this type of reasoning *monster-barring*, and gives examples from mathematics. Once the validity of a counterexample has been questioned, the focus of an argument switches from the *truth* of the conjecture to the *meaning* of its terms, which is negotiated by participants in a discussion according to their motivations and beliefs.

The interface between automated theory formation and situated embodied agents

A computational model of the embodied and social mathematics described above may also help to bridge the gap between automated theory formation and situated embodied agents. Despite forty years of research into automating the formation of mathematical theories, there is still no automated theory formation system which works at the pre-axiomatic stage or takes cognitively plausible knowledge as input. Conversely, although the subsumption architecture framework proposed by Brooks has proven itself in allowing the creation of reactive robots that can deal with the natural complexity of the real world, the architecture has proved somewhat limited in the complexity of the tasks to which it can be applied.

To allow robots, or embodied agents, to undertake more complex tasks, a return has been seen to the older sense-model-plan-act approach but with the robustness to the natural world being being built in at the modelling level through the use of powerful statistical techniques (Thrun, 2002). Recent work has proposed approaches which can build up concepts and rules about the world based on experience gained from interacting with a stochastic domain (Pasula et al., 2006; Shanahan, 2005). Being able to reason at a high level about these rules and concepts would be a powerful tool for an embodied agent learning about its environment, especially if such reasoning resulted in testable hypotheses that the embodied agent could try out in its world. Grounding a system of mathematics via embodied interaction with an environment would also relate to the symbol grounding problem; enabling us to provide an account of how mathematical language acquires meaning, and what this meaning might be.

A computational model of mathematical language acquisition and development

We are currently drawing from these ideas to produce a computational model of mathematical language acquisition and development. Such a model must comprise both an embodied level where mathematical ideas can be seen as hidden rules which hold for, or are inspired by, a physical world (based on Lakoff and Núñez's work), and an abstract level where these ideas are explored and sometimes changed (based on Lakatos's theory). We have already developed a computational model of Lakatos's theory and used our model to evaluate his theory (Pease et al., 2002; Pease et al., 2004). We envisage a 4-stage model in which the interaction between the embodied agent and the reasoning software would work in a simple arithmetic domain as described below.

An embodied agent is equipped with innate arithmetic capabilities such as ability to distinguish small numbers, subitizing, and perception of simple arithmetic relationships, as well as cognitive capacities including grouping, ordering, pairing, memory and metaphorizing (see (Lakoff and Núñez, 2001, pp. 51-52)), as well as ability to select and abstract common properties (see (Liebeck, 1984)). In the first stage the agent is able to interact with its environment, for example, by moving objects around into different piles and configurations, and to abstract properties of the group, such as its size. The agent may remember, or store, the results of adding a first pile to a second pile, and the results of adding the second pile to the first. This embodied interaction would lead to a set of concepts and facts about the environment which would then be passed as input to a theory formation system (which can be achieved with methods similar to those proposed by (Pasula et al., 2006; Shanahan, 2005) as described above). In the second stage this theory formation system would abstract and generalise rules which are descriptive of the patterns it finds. For example, it might generate the *commutative axiom of addition* (for natural numbers a and b, $a+b=b+a$). The system would then

explore the search space which the axioms define, by generating further concepts, making conjectures empirically, such as *whenever we subtract 1 from a number then we get another number*, and *all numbers can be written as the sum of two numbers*, and passing these to a theorem prover. In stage three, conjectures and theorems would then be passed back to the embodied agent for evaluation. For instance, the agent might evaluate relevance by testing whether a theorem can be instantiated within the world, or interestingness in terms of whether the theorem provides a new description of known behaviour or describes previously unknown behaviour. The agent might note that the two conjectures above hold for every collection of objects except for the collection of one object. It might then extend its concept of collection to including the empty collection, by performing the operation of removing one object from a collection of just that object and labelling the result a collection. Finally, in stage four, the same theory formation program would be used to analyse the information about the theorems and axioms and used to modify the axiom set. If one axiom had only been used to generate uninteresting theorems then this may be rejected at this stage. Conversely, for instance, having the "number" zero in the system might suggest further conjectures which would justify its inclusion in the theory. If any of the theorems contradicted each other then the axioms used would need to be modified or rejected.

We would evaluate our model based on whether it could reinvent concepts such as "zero" or axioms in a cognitively plausible way, and whether it recognised the interestingness of such pivotal concepts.

Conclusion

The theories we discuss in embodied and social mathematics are early characterisations of ways in which people do mathematics. We hope to build a computational representation of the theories which, starting from cognitively plausible innate abilities, models how we interact with an environment and how we formulate, explore, evaluate and modify axioms which describe that world. Our goal is to both extend and evaluate the theories we have discussed. It will be particularly exciting to further investigate the role that embodied interaction with an environment plays in our human mathematical development.

References

Butterworth, B. (1999). *What Counts: How Every Brain is Hardwired for Math*. Free Press, New York.

Choat, E. (1980). *Mathematics and the primary school curriculum*. National Foundation for Educational Research, Windsor.

Dienes, Z. P. (1973). *Six stages in the acquisition of mathematics*. Routledge.

Lakatos, I. (1976). *Proofs and Refutations*. CUP, Cambridge, UK.

Lakoff, G. and Núñez, R. (2001). *Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being*. Basic Books Inc., U.S.A.

Liebeck, P. (1984). *How Children Learn Mathematics*. Penguin Books, Middlesex.

Pasula, H. M., Zettlemoyer, L. S., and Pack Kaelbling, L. (2006). Learning symbolic models of stochastic domains. *J. of AI Research*, 29:309–352.

Pease, A., Colton, S., Smaill, A., and Lee, J. (2002). Semantic negotiation: Modelling ambiguity in dialogue. In *Proceedings of Edilog 2002, the 6th Workshop on the semantics and pragmatics of dialogue*.

Pease, A., Colton, S., Smaill, A., and Lee, J. (2004). A model of Lakatos's philosophy of mathematics. *Proceedings of Computing and Philosophy (ECAP)*.

Piaget, J. (2001). *Studies in Reflecting Abstraction*. Psychology Press, Hove, UK.

Shanahan, M. (2005). Perception as abduction: Turning sensor data into meaningful representation. *Cognitive Science*, 29:103–134.

Taylor, J. (1976). *The foundations of maths in the infant school*. Allen and Unwin, London.

Thrun, S. (2002). Probabilistic robotics. *Commun. ACM*, 45(3):52–57.

Ethical Implementation: A Challenge for Machine Ethics

Ryan S. Tonkens[1]

Abstract. The discipline of Machines Ethics, whose mandate is to create artificial moral agents (AMAs), is gaining momentum. Although it is often asked whether a given moral framework can be implemented into machines, it is never asked whether it should be. This paper articulates a pressing challenge for Machine Ethics: To identify an ethical framework that is both implementable into machines and whose tenets permit the creation of such AMAs in the first place. Without consistency between ethics and engineering, the resulting AMAs would not be genuine ethical robots, and hence the discipline of Machine Ethics would be a failure in this regard. Here this challenge is articulated through a critical analysis of the development of Kantian AMAs, which represents a leading contender for successful implementation of ethics into machines. In the end, the development of Kantian AMAs is found to be *anti*-Kantian. The upshot of all this is that machine ethicists need to look elsewhere for an ethic to implement into their machines.

1 INTRODUCTION

The nascent field of Machine Ethics is gaining momentum. Much of its fuel stems from the perceived imminent and inevitable[2] development of artificial moral agents (hereafter AMAs), who will be able to (or already *do*) perform morally consequential actions out in the world. Because autonomous machines will perform ethically relevant actions, akin to humans, prudence dictates that we design them to act morally.

Bracketed within the mandate of creating AMAs are issues regarding what sort of ethical framework robots ought to follow. For the most part, such concerns have rested on the question of how to *implement* a given moral framework into the machinery of the robot. Given this predominant focus on the engineering aspects of creating moral machines, competing ethical frameworks have been assessed based on their implementability, rather than the soundness of the theories themselves. One consequence of this is that it is never asked whether a given moral code *ought to be* implemented, only whether it *can be* done so successfully (in the sense that a genuine ethical robot would result[3]). In this light, finding the right ethic for machines to follow has come down to which one can best be implemented (from an engineering perspective), with other (ethical) issues falling by the wayside.[4]

Broadly speaking, this paper explores the issue of what sort of AMAs ought to be created. I take this question to be as fundamental as issues of how to best go about programming machines so as to be ethical, and they can each inform the other. For instance, if there is no way of *consistently* programming an AMA to follow a certain ethic, then perhaps such an AMA ought not to be built in the first place. Put differently, if our best options for implementing ethical frameworks into machines either cannot yield ethical machines, or if doing so goes against the tenets of those very same moral doctrines, then creating AMAs of that sort is morally dubious.

Achieving consistency between ethics and implementation represents a challenge for the field of Machine Ethics: To identify a moral framework that can be successfully implemented into machines, in such a way so that machines can (*do*) act ethically in the world, *and* whose own tenets permit the creation of AMAs in the first place. The bulk of this paper elucidates this challenge through an examination of Immanuel Kant's deontological moral framework.

Of the ethical doctrines being considered by machine ethicists, Kantian moral theory has become a frontrunner for putting the "ethic" into ethical machines. It is regarded by many as one of our best chances for the successful implementation of ethics into autonomous robots (See for example Anderson & Anderson 2006, 2007b; Powers 2006; Wallach et al. 2005). Although other frameworks have been proposed (See for example Grau 2006; Nadeau 2006; Allen et al. 2006), for my present purposes I assume that there is some weight to Anderson & Anderson's (2007a) claim that a duty-based approach is our most promising prospect in this vein. As the paradigmatic duty-based ethic, Kantian morality promises to offer an implementable moral framework for our robots to successfully abide by. Despite this possible *implementability*, what has yet to be asked is whether

[1] Dept. of Philosophy, York Univ., Toronto, Canada. Email: tonkens@yorku.ca.

[2] Allen et al. (2006), page 13. See also Sparrow (2007), page 64.

[3] By "genuine ethical robot" and variations thereof I am referring to both an autonomous machine that is able to (and does) act in a manner consistent with a given ethical framework, and also a robot that possesses the qualities necessary for moral agency.

[4] Although not attended to herein, one example here surrounds the issue of competing views regarding what the best moral theory is. For a proponent of *virtue ethics* (for example), it may be thought that there is only *one* moral framework that should be implemented into machines in order to render them moral (i.e. virtue ethics), since all other competing moral theories are deemed dissatisfactory in some way(s). *If* a virtue ethic could not be successfully implemented into machines, then the creation of AMAs would be impossible (according to the ardent virtue ethicist), since following a virtue ethical framework is the only way (for any*thing*) to act morally in the first place. Machine Ethics as a discipline has thus far ignored this (and other) *ethical* issue(s), since its primary goal is to create AMAs, irrespective of the ethical theory that needs to be appealed to in order to do so. (Thank you to the anonymous reviewer for bringing this point to my attention).

Kantian ethics permits the development of AMAs in the first place.[5]

Once this question is asked, it becomes clear that creating *Kantian* artificial moral agents is *anti*-Kantian. On one hand, Kantian moral machines would not be Kantian moral *agents*, strictly speaking. On the other hand, even if such machines were Kantian moral agents, their creation would nevertheless violate Kantian moral law. Because of this, the creation of Kantian AMAs is inconsistent with the prescriptions of Kantian morality. Since we (rightly) demand consistency in ethics, the failure of such machines to meet the standards of morality that they are designed to heed is unacceptable. The upshot of all this is that machine ethicists need to look elsewhere for an ethic to implement into their machines.

The course of this paper runs as follows. In section two I elaborate on what I take to be a serious challenge for Machine Ethics. Meeting this challenge is crucial for achieving the underlying goals of Machine Ethics in general. Section three represents an exposition of the sort of robot that is under issue. In section four I review the basic tenets of Kantian morality, with the goal of setting the stage for my critique of the creation of Kantian AMAs in section five. In the closing sections, I examine the scope and limitations of this paper, and conclude with some suggestions for future research in Machine Ethics.

2 A CHALLENGE FOR MACHINE ETHICS

Much of the work being done in Machine Ethics concerns issues of how to best implement a given moral framework into the machinery of a robot, so as to render it ethical. Many different proposals have been advanced, some of which are quite promising, at least from an *engineering* perspective. What has yet to come up is the idea that some of the ethics we are trying to implement into machines may not allow for the creation of AMAs in the first place. Although it is correct to ask "if ethics is the sort of thing that can be computed",[6] we also need to ask whether a given ethic *should be* computed. Moreover, Allen et al. (2006) are correct to suggest that machine ethicists "must assess any theory of what it means to be ethical or to make an ethical decision in light of *the feasibility of implementing the theory as a computer program*" (emphasis added, 15). But this does not demand enough; we must also assess our ethical theories in light of whether those theories allow for the development of AMAs, prior to the implementation stage. Demanding that our moral machines act differently than we do, or to permit violations against the same moral framework that we have programmed robots to obey, is *ethically inconsistent.* Meeting this consistency constraint represents a challenge for Machine Ethics.

Although we may come to expect our robots to be *more moral* than humans in some ways[7], the moral standing of human action is in some sense projected onto the very existence of the machine. Although humans may not be Kantians, and may even sometimes violate Kantian morality, by creating a Kantian AMA and demanding that it obey Kantian morality, we are asking it to achieve the impossible (as discussed below). Through its very creation the machine cannot be moral; the development of such AMAs is already an ethical breach. This is not to say that the creation of AMAs *in general* is not permitted, only that the development of *Kantian* AMAs contradicts *Kantian* ethics. This point is worth labouring: The view endorsed herein remains optimistic that the successful *and* ethical development of moral machines is possible. The point is that, in order to do so, we need to find a match between what ethical frameworks we can implement and those we are allowed to implement. Before giving flesh to these arguments, we need to know what kind of machine is under issue here.

3 ARTIFICIAL MORAL AGENCY

The sort of AMAs under issue herein are machines that can make decisions and perform actions in real world contexts (based on Kantian ethics), where such actions may have moral consequences. That such machines may come to fruition is undeniable (See for example Allen et al. 2006; Moor 2006; Wallach et al. 2008). According to Anderson & Anderson (2006, 2007a, 2007b), the ultimate goal of Machine Ethics is to create a machine that is an explicit ethical agent. Gips (1995, 2005) even goes so far as to suggest that the creation of ethical robots ought to be considered a Grand Challenge for computing research and AI. In light of this, I take the development of AMAs that can act in the world to be *the* main goal of Machine Ethics.

What exactly, then, is an artificial moral agent (or an ethical robot or a moral machine)? I follow Moor (2006) in distinguishing between four types of artificial moral agent: i) ethical-impact agents, ii) implicit ethical agents, iii) explicit ethical agents, and iv) full ethical agents. According to Moor, *ethical-impact agents* are those computing technologies that have ethical impacts on their environment in some way. Moor offers the example of contemporary camel racing practices in Qatar, where human slave-boys have been replaced with robots as the camel jockeys, thus relieving the boys of a life of forced servitude. Another example is the atomic bomb (as a weapon of mass destruction). There is really no *agency* at this level, nor is there any sign of self-directed action either. These machines are ethical agents in the very weak sense that their functions serve purposes that have moral consequences (whether directly or indirectly).

A step up from these impact agents is what Moor calls *implicit ethical agents*, which are machines that are designed to implicitly follow some sort of ethical rule. I take Moor to be suggesting that such machines could not act immorally, mostly because they could not really *act* in any strong sense in the first place. Automatic pilots in aircrafts and automated bank tellers are examples of this sort of machine. The very design of these 'agents' implicitly constrains their behaviour to morally acceptable actions. The important points to highlight here are that i) these machines cannot act counter-ethically or 'immorally' and ii) although they do 'act' out in the real world, there is little to no agency (autonomy) at this level. If the machine 'acts' wrongly (as a result of its malfunctioning perhaps), the designer or the user is to blame, rather than the

[5] To my knowledge, the most thorough examination of Kantian ethics within the Machine Ethics literature thus far is offered by Powers (2006). Although such an analysis of Kant's ethics is a step in the right direction, Powers' discussion all along remains at the level of *implementation*, and never considers whether Kantian morality permits the development of Kantian AMAs in the first place.

[6] Anderson & Anderson (2007b), page 18.

[7] Nadeau (2006) even goes so far as to suggest that *only* androids could be ethical.

machine itself. In this way, as Moor rightfully points out, a machine's 'capability to be an implicit ethical agent doesn't demonstrate their ability to be full-fledged ethical agents' (19).

Agents falling into the next two categories have the distinctive feature that, not only can they (often) act out in the world, but they can do so with little *to no* human supervision. *Explicit ethical agents* have the ability to make explicit ethical judgments and to justify them. Examples here include autonomous automated military weapons currently being proposed (or already in use), mostly in the United States.[8] One particularly interesting example is the latest Unmanned Underwater Vehicle (UUV), labeled MANTA, which is presently being researched by the U.S. Navy. This machine will be 'capable of autonomously seeking out, attacking, and destroying enemy submarines' (Sparrow 2007, 63). Explicit ethical agency is best understood juxtaposed with Moor's characterization of *full ethical agents.* Full ethical agents go beyond explicit ethical agency since they also possess capacities such as (self-)consciousness, intentionality, emotion, creativity, free will, *et cetera*.[9] The paradigmatic full ethical agent is a 'normal' adult human being. To my knowledge, at the time of writing this paper, no machine has reached the status of being an authentic full ethical agent. Much of what makes Machine Ethics relevant is that it investigates whether it is possible to do so, and helps to prepare just in case it is.

It is worth noting that debate continues over the notion of machine agency. Some have argued that machines cannot be (moral) agents in any significant sense of the term. Johnson (2007), for example, argues that computer systems may be moral *entities*, although they cannot be moral *agents*. Sparrow (2007) has argued that, although machines may be autonomous, they cannot be held morally responsible for their actions. Torrance (2008) argues that AMAs would not be authentic members of the moral community since they would lack certain crucial characteristics unique to biological entities. On the other hand, it has been argued that machines can in fact be (full) ethical agents, although perhaps only when understood at a certain level of abstraction (Floridi & Sanders 2007). Some have gone so far as to argue that robots could (should) be afforded the legal status akin to persons (Calverley 2008). I do not have much to say by way of commenting on this debate here. The important point for our purposes is not whether AMAs can meet the criteria for *any* characterization of moral agency, but whether they could meet the criteria for *Kantian* moral agency. This is because I am not arguing against the creation of AMAs *in principle*, but rather that the creation of *Kantian* moral agents violates Kantian ethics. If our (non-Kantian) AMAs turn out to meet different standards for moral agency, then so much the better for Machine Ethics.

What I am concerned with in this paper is any machine that falls into the categories of *explicit* or *full* ethical agent. Of the characteristics possessed by these more developed forms of artificial agent, the capacity for self-directed *action* out in the world is particularly germane to our discussion. Were our ethical machines to remain barred from acting out in the world, then many of the worries mounted herein are misplaced (I return to this point in §6). Equally, much of what is at stake here rests on whether our ethical robots will be *autonomous* to any significant extent. As is argued for below (§5), if such robots are *not* autonomous, then they are not Kantian agents, and hence they would not be (*could* not be) consistently bound by Kantian morality. On the other hand, if they *are* autonomous, then, although they may be Kantian moral agents, their existence nevertheless represents a moral breach. Before making these arguments, a brief exposition of Kantian ethics is necessary.

4 KANTIAN ETHICS

For our purposes, two main ideas of Kantian ethics need to be highlighted: i) the foundations of moral agency and ii) the role of the categorical imperative in moral decision making.[10] Each is taken up in turn.

According to Kant, moral agency has two overarching components: *rationality* and personal freedom (or *autonomy*). Only those beings that are rational and free are (or can be) moral agents. The moral law stems from pure reason alone, outside of experience (*a priori*), and is necessarily and universally binding on all rational beings as such. The objective law of morality, as a law of reason, acts as a compass for moral action. Human volition, as the willing of a subject that is both rational and sensible, is necessarily faced with cases of conflict between these two competing natures. Whereas inclination serves to secure pleasure and the basic needs for survival (in short, contingent means to largely animalistic ends), reason has a different role to play. Reason guides action in accordance with objective laws, towards the end of establishing a good will and moral character.

The competing natures of human beings will come up again later on (§5). According to Kant, it is only because humans *can* violate the moral law and succumb to the temptations of sensual satisfaction that they can truly be said to be moral agents. Duty signifies the (rational) "strength needed to subdue the vice-breeding inclinations".[11] In other words, part of the force and achievement of acting dutifully stems from the fact that *one could have acted otherwise* (i.e. *non-dutifully*).

Moral agents can act contrary to duty (albeit *immorally*) since they possess free will. Freedom has both a negative and a positive conception, according to Kant. A moral agent is free in a *negative* sense insofar as no foreign causal forces dictate what she, as a rational agent, ought to do. Moral agents are free in a *positive* sense insofar as reason is freely able to give to itself and follow laws of its own fabrication—the will as subject only to its own laws. Moral agents are thus *fully autonomous* and independently lawmaking beings. According to Kant, 'the idea of morality reduces to the idea of freedom'; we are driven to presuppose the concept of freedom in order to understand ourselves as initiating moral causation, and hence as conceiving all rational beings as exhibiting such causation.[12] In this way, the categorical *ought* reveals itself as reason's tool for *rational self-determination* in the face of inclinational temptation.

With rationality and freedom as the two points of departure for morality, Kant proceeds to articulate the moral law through

[8] For a nice review of these weapons, see Sparrow (2007).

[9] For a recent interdisciplinary discussion of creativity, see Boden (ed.) (1994). For a discussion of the intersection of emotions and AI, see Picard (1997).

[10] For a more in depth analysis of Kant's moral philosophy, see O'Neill (1989) and Rawls (2000).

[11] *The Metaphysics of Morals*, page 141. (Hereafter MM).

[12] *Fundamental Principles of the Metaphysic of Morals*, page 80. (Hereafter FPMM).

the conception of what he terms *the categorical imperative*. In order to assess whether an action is morally permissible or not, an agent must test her subjective maxim—her personal principle of action—against the objective formal criteria of the categorical imperative. In order for acting upon a maxim to be moral, that maxim needs to be *universalizable*. Roughly, it must be consistently held that all moral agents, given the same context, would (*could*) act on that very same maxim. The categorical imperative can be seen as a heuristic for determining what actions are dutiful and which ones are not. For our purposes, the first two formulations of the categorical imperative are worth making explicit:

CI-1: Act only on those maxims whereby you can at the same time will that they should become universal laws.[13]

CI-2: So act as to treat humanity [i.e. moral agency], whether in your own person or in that of any other, in every case as an end in itself, and never merely as a means.[14]

Later on (§5), the maxim that sanctions the development of Kantian AMAs is applied to the categorical imperatives stated above. Through doing so, we can better understand the moral implications of developing this type of ethical robot.

Kant's moral framework is *deontological*, meaning that it is founded on the idea that doing what is right is none other than doing one's *duty*. According to Kant, rational beings determine their duties for themselves, through exercising their rationality. Acting dutifully is the only path towards establishing a good will, which is the only thing that is good without qualification.[15] In order for an action to be moral, it must both conform to *and stem from* duty. In cases where actions are not done for the sake of duty (for example, actions that are committed through reflex), despite the fact that they may *conform to* duty, are not moral, strictly speaking.

This is Kantian ethics in a nutshell. According to Kant, moral actions are those that conform to the categorical imperative (the objective law of morality), are done out of duty (for morality's sake), and are committed by beings who are rational and free (moral agents). In what follows, it is argued that artificial moral agents *cannot be Kantian*. This is the case since they would not be free, and since their creation violates the categorical imperative in several ways. For these reasons, implementing a Kantian ethic into robots has already gone too far. In this way, adopting a Kantian perspective towards creating ethical machines does not meet the challenge noted above, namely, to identify an ethic that we are *ethically permitted* to implement.

5 KANTIAN AMAs ARE *ANTI*-KANTIAN

In this section, it is argued that the creation of Kantian artificial moral agents is not consistent with Kantian ethics. This is because Kantian AMAs would not be Kantian moral *agents*, and since the creation of Kantian AMAs violates the categorical imperative in several ways. Because we require our Kantian AMAs to act ethically, the fact that their development is a violation of Kantian morality renders their creation morally suspect, and our role as their creators hypocritical. The upshot of this is that, despite the idea that Kantian ethics *may* be implementable into machines, these types of machines should not be developed, at least to the point where they are able to act out in the world. I offer three arguments to support these claims.

1. *Kantian AMAs would not possess free will.* Recall that the nature of moral agency, according to Kant, is twofold: Moral agents are both rational and free. In cases where one or both of these attributes are absent, then genuine moral agency is absent as well. Here I assume that AMAs will (eventually) be rational. If this assumption turns out to be misguided, then so much the better for my argument as a whole; without rationality, AMAs would not be Kantian moral agents. Be this as it may, what interests me here is whether or not AMAs would be free, so as to satisfy both requirements for Kantian moral agency.

The extent to which AMAs would be *programmed* to act in certain specific ways seems to prevent their being free. In fact, *all* of the machine's actions would be predetermined by the rules that it was programmed to follow.[16] Beings that are determined in all of their actions do not possess free will. This is especially evident in the fact that machine ethicists are going to such lengths to make sure that machines act ethically in the first place; the goal of Machine Ethics is to create an ethical robot, not one who *sometimes* acts ethically, or that can act *un*ethically.[17]

On one hand, AMAs would be programmed to act according to the rules that the programmer has installed in them. The resulting AMA is in this way determined to perform certain functions, to act in certain ways, and to hold certain epistemic truths, *et cetera*. What is more, it could not act otherwise than how it has been programmed to act (assuming optimal functioning). On the other hand, AMAs would also be programmed to *not* perform certain functions, not act in certain ways, *et cetera*, despite their otherwise *having the potential for* doing so. In this way, Kantian AMAs would not be free in both the positive and negative sense of freedom outlined above. Kantian AMAs are not free in the positive sense since the rules of the programmer (*and not of the machine itself*) constrain the machine's actions. In the same way, to the extent that the intentions of the programmer represent *foreign* causal forces dictating its actions, the AMA is not free in a negative sense either. The rules that AMAs follow are given to them from the exterior, and hence they are not of their own making.

For example, that the 'killer robot' used for military purposes *could not withhold* gunfire when it is given *sound* orders to open fire demonstrates its lack of freedom. It is important to note that withholding assault would not be a *moral* violation (at least in most cases). This is important since it is not merely by *being ethical* that AMAs would *necessarily* not be able to perform certain actions, and hence have reduced (or non-existent) freedom. Rather, withholding assault could not occur because the AMA has not been programmed in such a way so as to allow

[13] FPMM, page 49.

[14] FPMM, page 58.

[15] FPMM, pp. 17-20.

[16] Admittedly, this claim is controversial. Some have argued that free will *could* be instilled in robots. See especially John McCarthy's "Free Will—Even for Robots" (2000).

[17] Although not discussed herein, perhaps a *compatibilist* understanding of free will may hold some promise here. Yet, to the extent that Kant's view is properly positioned in the *indeterminist* camp, and to the extent that we want our robots to be *determined* to only act ethically, then a compatibilist approach would require making sacrifices that a *Kantian machine ethicist* may be reluctant to condone.

for the voluntary dismissal of sound commands. Even as the machine learned to apply such rules to novel cases, it would never have (unfettered) control over its actions. The point is that an AMA could only act from within the given domain of those actions that are in conformity with the rules manifested in its machinery *by its programmer*. So, the military AMA could withhold fire in certain contexts (when the targets in sight are innocent civilians, say), but it could never resist its programming to follow sound orders (to open fire on legitimate enemy targets). Furthermore, *we may not want* our AMAs to be able to act freely, especially to the extent that this could result in unethical behaviour on their part. Surely not all actions done from free will represent moral violations. But, in the case of AMAs, protecting against cases of ethical violations means prohibiting certain actions from being able to be done in the first place. In this way, Allen et al. (2000) are correct to suggest that "human-like performance, which is prone to include immoral actions, may not be acceptable in machines" (251).

In addition to this, regardless of whether we would want our AMAs to possess freedom of the will (and there is reason to think that we would not), in order for them to be *Kantian* moral agents, they would need to be free (and rational). In fact, they would need to be free to the extent that their actions were only genuinely moral since they marked an autonomous overcoming of non-dutiful inclination—otherwise, their actions would not be bound by the moral law, nor could they be held responsible for their actions. According to Kant, part of being a moral agent means possessing "the capacity to master one's inclinations when they rebel against the [moral] law", hence the ability to freely commit actions that are not moral.[18] The goal of Machine Ethics, however, is precisely to reduce morality in robots to something like *unchallengeable inclination*.

On one hand, then, if (Kantian) AMAs are not free, then they are not authentic Kantian moral agents. In this case, the machine would not (*could not*) come to view itself as being bound by the moral law, and our goal of creating *ethical* machines would be a failure in this regard. Although the AMA would most likely act morally—for it wouldn't have the freedom to do otherwise—it would nevertheless lack moral agency. What is more, since the AMA is not a proper moral agent, then neither is it the proper target of praise or blame. (This last point will be discussed in greater detail below). On the other hand, if AMAs *are* free, then they would be able to willfully act immorally (if they so choose to), regardless of what their programming dictates. Allen et al. (2000) recognize this point when they write:

> If, as Kant appears to think, being a moral agent carries with it the need to *try* to be good, and thus the capacity for moral failure, then we will not have constructed a true artificial *moral* agent if we make it incapable of acting immorally. Some kind of autonomy, carrying with it the capacity for failure, may be essential to being a real *moral* agent. (Original emphasis).[19]

The only way to develop authentic *Kantian* moral agents would be to create AMAs that are free, at least to the extent that they have the choice of whether to act morally or to act immorally. This is most likely not a consequence that machine ethicists would be willing to accept, and rightly so.

Even if a case can be made that AMAs could be Kantian moral agents, who are both rational *and* free, their creation nevertheless violates Kantian ethics in other ways. The remaining arguments all surround the idea that the development of Kantian AMAs violates the categorical imperative in some way. Because of this, it is helpful to make explicit the subjective maxim that the developer of Kantian AMAs might propose to universalize. The Machine Ethics Maxim (MEM) may be articulated as follows:

MEM: So act as to will the creation of autonomous *Kantian* explicit (or full) artificial moral agents that can perform morally consequential actions out in the world.

MEM fails to uphold the (Kantian) moral law in several ways. This is not to say that different ways of formulating it may not avoid this outcome. For example, were we to replace "Kantian" with "Virtuous" (say), a separate investigation would be needed to assess whether creating such AMAs is consistent with the tenets of Virtue Ethics. Moreover, none of this is to say that AMAs ought not to be created at all, ever. The important point is this: Part of the requirements for successfully implementing a moral framework into robots is for that moral doctrine to allow for the creation of that type of AMA in the first place. Although Kantian ethics *may* be implementable, doing so contradicts the tenets of Kantian ethics. It remains an open question whether other moral codes may fare any better.

2. *The creation of Kantian AMAs violates CI-2*: By creating Kantian moral machines, we are treating them merely as means, and not also as ends in themselves. According to Kant, moral agents are ends in themselves and they ought to be respected as such. To violate this law is to treat an agent merely as an *object*, as something *used for* achieving other ends.

It is unclear whether machines could be treated as ends in themselves in the first place. According to Kant:

> [A] human being [*qua* moral agent] regarded as a *person*, that is, as the subject of a morally practical reason, is exalted above any price; for as a person (*homo noumenon*) he is not to be valued merely as a means to the ends of others or even to his own ends, but as an end in himself, that is, he possesses a *dignity* (an absolute inner worth) by which he exacts *respect* for himself from all other rational beings in the world. He can measure himself with every other being of this kind and value himself on a footing of equality with them. (Original emphasis).[20]

In order to be treated as an end in itself, a Kantian AMA would need to possess dignity, be deserving of respect by all human beings (all other moral agents), and be valued as an *equal* member in the moral community. Such equality entails personal rights, opportunities, and status akin to that of human beings

[18] MM, page 148.

[19] Allen et al. (2000), page 254.

[20] MM, page 186.

(among other things, perhaps[21]). The default position here should be to refrain from granting *equal* rights, opportunities, and status to machines, at least until AMAs become sophisticated enough so as to be widely (and uncontroversially) recognized as being genuine full moral agents. At any rate, the burden is on those who want to afford equal rights (both human and moral) to machines to offer good reasons for doing so.

Regardless, as the present state of the art indicates, humans have no intentions of treating ethical robots as anything other than means to (*anthropocentric*) ends. This becomes obvious once we examine some of the reasons typically advanced for creating AMAs in the first place. Allen et al. (2000) suggest that robots "possessing autonomous capacities to do things that are *useful to humans* will also have the capacity to do things that are *harmful to humans* (emphasis added, 251). Moor (2006) summarizes three general reasons in favour of developing explicit ethical machines:

1. Ethics is important. We want machines to treat us well.

2. Because machines are becoming more sophisticated and make our lives more enjoyable, future machines will likely have increased control and autonomy to do this. More powerful machines need more powerful machine ethics.

3. Programming or teaching a machine to act ethically will help us better understand ethics.[22]

I take Moor's reasons for creating AMAs to be fair enough. If machines were able to treat human beings in any morally relevant manner at all, then we would want them to treat us well. Equally, it seems correct to suggest that, were machines to be powerful agents in the world, then we would want them to be equally as ethical. Moreover, Moor is not alone in arguing that research in Machine Ethics may be insightful with respect to understanding ethics as a whole. As Anderson & Anderson (2006) have put it, "machine ethics, by making ethics more precise than it's ever been before, could lead to the discovery of problems with current ethical theories, advancing our thinking about ethics in general".[23]

Despite their reasonableness, these reasons are all oriented towards the satisfaction of human ends—the protection of humans from ethical wrongdoing, the improvement of human understanding of morality, robots as ethical advisors to humans, and the creation of machines for increasing human enjoyment, *et cetera*—and pay no attention to the machine as an end in itself. Because of this, as reflected in the current state of the art, the creation of Kantian AMAs seems to violate the second formulation of the categorical imperative. In this way, the development of Kantian AMAs in *anti*-Kantian.

In his "Towards the Ethical Robot" (1995), Gips argues that "the robotic/AI approach…tries to build ethical reasoning systems and ethical robots *for their own sake*, for the possible benefits of having the systems around as actors in the world and as advisors, and to try to increase our understanding of ethics" (emphasis added, 11). Worth noting is that most of Gips' reasons here are anthropocentric (just like those noted above). The interesting idea that Gips suggests is that ethical robots could be built "for their own sake". If this is true, then perhaps such AMAs would be (could be) treated as ends in themselves, rather than merely as means. If this is the case, then the creation of Kantian AMAs may be consistent with CI-2 after all.

But Gips does not offer any reason to back up his claim here. In fact, it is difficult to see how *building* ethical machines could be done for their own sake, even if we wanted to do so. For one thing, prior to their creation, there is no "sake" for them to have.[24] Furthermore, even if it is assumed that we may *in principle* be able to treat *existing* AMAs as ends in themselves, doing so is only possible if we drastically reorient our *reasons for wanting to create them in the first place*, placing greater (primary) emphasis on non-anthropocentric (*robocentric*) ends, as opposed to the anthropocentric ones currently on offer. Being charitable to Gips, perhaps there is a way that ethical machines could be (treated as) ends in themselves or could be created for their own sake. However, the burden of proof is on him to support this claim. In the absence of such support, the creation of Kantian AMAs is a violation of CI-2.

3. In light of what has been said thus far, *the creation of Kantian AMAs is a violation of CI-1 as well*: By creating Kantian AMAs, we would be implying their inclusion into the group made up of all other moral agents (human and android alike). In fact, the only way it would work is if such robots were subject to moral praise and punishment (Sparrow 2007). Because of this, when testing their maxims (e.g. MEM), (human) agents would need to consider AMAs as being included in the pool of agents for whom that maxim could be universalized. Yet, because we would be treating AMAs *merely as means to human ends* from the beginning, AMAs themselves would be forced to not condone MEM (as an *ongoing* maxim), since they would understand it as being inconsistent with the (Kantian) moral code that they have been programmed to obey. Kantian AMAs would recognize MEM as *non*-universalizable, since it entails the violation of CI-2 (as discussed above), and hence as not being a maxim that could be acted upon by *all* moral agents (consequently violating CI-1). AMAs would not condone their being treated merely as means, and hence would not endorse MEM, consequently rendering MEM a violation of CI-1.

It is worth noting that this problematic outcome cannot be avoided simply by omitting to include AMAs as members of the wider group of moral agents during the process of moral deliberation, since their genuine membership in this group is crucial for their being bound by Kantian moral law. If such machines were not bound by moral law, then they would not be ethical machines, strictly speaking. In perhaps the worst case scenario, such robots would understand their very existence as not being consistent with the moral code that they were designed to follow, and hence may come to understand their existence as being something morally *abhorrent*. In such (admittedly

[21] For instance, we *may* be reluctant to afford a *non-conscious* or a *non-sentient* AMA rights akin to human beings. For an interesting discussion in this vein, see Torrance (2008).

[22] J. Moor (2006), page 21.

[23] Anderson & Anderson (2006), page 11.

[24] Even the idea of "potential sake" seems strange in this context; through the process of building an AMA, we would be simultaneously creating the machine *and* any sake that the machine may come to have. That the machine may come to condone its having been created (in the sense that it may be grateful for having been given the opportunity to be an end in itself and to have a sake) says nothing about *why* and for *what reasons* the AMA was created in the first place.

speculative) instances, we may find AMAs in a state of moral paralysis or existential alienation. We may even find our ethical robots turning to (what Kant called) *heroic suicide* in order to preserve morality in the world.[25] If Kantian AMAs were *not* authentic moral agents, then none of this would occur. This, however, would mean that they were not the sort of ethical robot that machine ethicists are aspiring to build, and would come at the expense of not being able to hold such AMAs morally responsible for their actions. If they *were* Kantian moral agents, then their being programmed to abide by the moral law commands them to recognize their existence as inconsistent with Kantian morality, since it violates both CI-1 and CI-2.

In this section I have offered three arguments to suggest that the creation of Kantian AMAs is inconsistent with Kantian ethics. The main upshot of all this is that machine ethicists need to look elsewhere in search of a moral code to implement into autonomous machines.

6 THE SCOPE AND LIMITS OF THIS PAPER

As noted earlier, this critique does not apply to the creation of all moral machines. All of the arguments mounted against the development of Kantian AMAs surround the idea of their (not) being authentic moral agents who *act out in the world*. If we restrict the role of Kantian machines so that they do not act in the world, perhaps to that of an advisor to humans in making moral decisions, then the worries mounted herein readily dissolve.

Examples of such machines include MEDETHEX, a machine devised for giving bioethical advice (Anderson & Anderson 2007b), McLaren's (2006) TRUTH TELLER, which is a "computational model of casuistic reasoning" designed to help students discriminate between cases of truth-telling and lying, and the connectionist network designed by Guarini (2006) that can successfully apply moral rules that it has learned to novel cases. Machines such as these promise to fulfill the goal of using machines to help humans better understand ethics as a whole. None of these machines can act out in the world, and hence none of their actions could have (direct) moral consequences. These machines are not taken to be genuine autonomous moral agents, and hence their existence does not require that our implementation practices be consistent with the ethical frameworks they are designed to follow. Once our robots move out into the world, however, then ethical consistency becomes indispensable.

Although this paper is largely critical in nature, it has a positive implication for Machine Ethics as well. By demonstrating that a Kantian AMA is a contradiction in terms, our pool of possible ethical frameworks for successful implementation into machines is consequently narrowed.[26] In this way, we are closer to finding the proper ethic for implementing into machines than before. This point serves to emphasize the idea that the self-imposed ultimate goal of Machine Ethics—to create autonomous ethical robots that act out in the world—is not necessarily something that is morally impermissible through and through.

Nevertheless, *all* moral frameworks that are considered for implementation into machines need to be assessed with respect to whether they permit the development of AMAs, prior to the implementation stage.[27] The challenge for Machine Ethics proposed here is to maintain consistency between what we want to implement and what we ought to implement. Finding a moral framework that meets these demands is certainly not impossible *in principle*. Future research in this area should therefore not be restricted to issues of implementation. Researchers should also consider the *ethical* dimensions of choosing a framework for eventual implementation. Otherwise, the goal of creating authentic ethical machines is significantly threatened.

7 CONCLUDING REMARKS

> I wish to build completely autonomous mobile agents that co-exist in the world with humans, and are seen by those humans as intelligent beings in their own right […] I have no particular interest in applications; it seems clear to me that if my goals can be met then the range of applications for such Creatures will be limited only by our (or their) imagination. I have no particular interest in the philosophical implications of Creatures, although clearly there will be significant implications.[28]

The burden of this paper has been to explore some of the philosophical implications of creating Kantian artificial moral agents ('Creatures'). At least with respect to Kantian ethics, AMAs that can act in the world ought not to be created. It was argued that this is the case since Kantian machines would not be Kantian moral agents, and hence would not be bound by Kantian moral law. Furthermore, even if Kantian AMAs could be authentic moral agents, their very existence violates (at least) the first two formulations of the categorical imperative.

Our machine ethic needs to be consistent in the sense that the moral framework being implemented into our machines allows for the development of such artificial moral agents in the first place. Where this consistency is absent, our robots will not be genuine ethical agents, and their developers would hypocritically demand that such robots conform to a doctrine that they themselves violated during the act of creation. The worry is that putting AMAs into the world without first establishing such a consistency is ethically dubious. Kantian moral machines are *non*-Kantian, and hence fail to establish this required consistency. This remains the case despite the *possibility* of successfully implementing Kantian ethics into machines. The upshot of all of this is that we need to find a better candidate for

[25] See Kant's *Lectures on Ethics*. There Kant distinguishes between *heroic* (supererogatory), *blameworthy* (abhorrent), and *permissible* (accidental) suicide. Heroic suicide represents self-termination that is done with the intent of maintaining morality in the world, most notably in cases where remaining alive would initiate a more severe moral violation.

[26] There may be some Kantian machine ethicists that would rather give up on Machine Ethics altogether than resort to implementing a non-Kantian moral theory into AMAs (insofar as they are convinced by the arguments presented herein). Yet, for those whose paramount goal it is to create AMAs (regardless of what ethical framework they need to appeal to in order to do so successfully), the upshot of this paper should spark a redirection in focus rather than the forfeiture of that goal.

[27] It is worth noting that several authors have recognized certain difficulties that surround the *implementation* of Kantian ethics into machines. See for example Wallach et al. (2008), Allen et al. (2000, 2005), and Gips (1995).

[28] R. Brooks (1991), page 145.

an ethic that is both implementable, *and* whose tenets permit the creation of AMAs in the first place. I consider this to be a serious challenge for the discipline of Machine Ethics.[29]

REFERENCES

[1] Allen, C., W. Wallach, & I. Smit (2006): Why Machine Ethics? *IEEE Intelligent Systems*, 21 (4): 12-17.

[2] Allen, C., I. Smit, & W. Wallach (2005): Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. *Ethics and Information Technology*, 7: 149-155.

[3] Allen, C., G. Varner, & J. Zinser (2000): Prolegomena to Any Future Artificial Moral Agent. *Journal of Experimental and Theoretical Artificial Intelligence*, 12 (3), 251-261.

[4] Anderson, M. & S. L. Anderson (2007a): The Status of Machine Ethics: A Report from the AAAI Symposium. *Minds and Machines*, 17: 1-10.

[5] Anderson, M. & S. L. Anderson (2007b): Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine*, 28 (4) : 15-27.

[6] Anderson, M. & S. L. Anderson (2006). Machine Ethics. *IEEE Intelligent Systems*, 21 (4): 10-11.

[7] Anderson, M., S. L. Anderson & C. Armen (2006): An Approach to Computing Ethics. *IEEE Intelligent Systems*, 21 (4): 56-63.

[8] Boden, M. A. (ed.) (1994): *Dimensions of Creativity*. Cambridge: MIT Press.

[9] Brooks, R. A. (1991): Intelligence without Representation. *Artificial Intelligence*, 47: 139–159.

[10] Calverley, D. J. (2008): Imagining a Non-biological Machine as a Legal Person. *AI & Society*, 22 (4): 523-537.

[11] Floridi, L. & J. W. Sanders (2004): On the Morality of Artificial Agents. *Minds and Machines*, 14 (3): 349-379.

[12] Gips, J. (2005): Creating Ethical Robots: A Grand Challenge. *AAAI Symposium on Machine Ethics*, Washington, D.C.

[13] Gips, J. (1995): Towards the Ethical Robot. In K. Ford, C. Glymour, & P. Hayes (eds.), *Android Epistemology*. Cambridge: MIT Press, 243–252.

[14] Grau, C. (2006): There is no 'I' in 'Robot': Robots and Utilitarianism". *IEEE Intelligent Systems*, 21 (4): 52-55.

[15] Guarini, M. (2006): Particularism and the Classification and Reclassification of Moral Cases. *IEEE Intelligent Systems*, 21 (4), 22-28.

[16] Johnson, D. G. (2006): Computer Systems: Moral Entities but not Moral Agents. *Ethics and Information Technology*, 8: 195-204.

[17] Kant, I. (1997): *Lectures on Ethics*. Trans. P. Heath. Cambridge: Cambridge University Press.

[18] Kant, I. (1996): *The Metaphysics of Morals*. Trans. M. Gregor. Cambridge: Cambridge University Press.

[19] Kant, I. (1988): *Fundamental Principles of the Metaphysic of Morals*. Trans. T. K. Abbott. New York: Prometheus.

[20] McCarthy, J. (2000): Free Will—Even for Robots. *Journal of Experimental and Theoretical Artificial Intelligence*, 12(3): 341–352.

[21] McLaren, B. (2006): Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions. *IEEE Intelligent Systems*, 21 (4), 29-37.

[22] Moor, J. H. (2006): The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, 21 (4), 18-21.

[23] Nadeau, J. E. (2006): Only Androids Can be Ethical. In K. Ford, C. Glymour, & P. J. Hayes (eds.), *Thinking About Android Epistemology*. Cambridge: MIT Press, pp. 241-248.

[24] O'Neill, O. (1989): *Constructions of Reason: Explorations of Kant's Practical Philosophy*. New York: Cambridge University Press.

[25] Picard, R. W. (1997): *Affective Computing*. Cambridge: MIT Press.

[26] Powers, T. M. (2006): Prospects for a Kantian Machine. *IEEE Intelligent Systems*, 21 (4): 46-51.

[27] Rawls, J. (2000): *Lectures on the History of Moral Philosophy*. Cambridge: Harvard University Press.

[28] Sparrow, R. (2007): Killer Robots. *Journal of Applied Philosophy*, 24 (1): 62-77.

[29] The U. S. Army Future Combat Systems Program (2006). Available at www.cbo.gov/ftpdoc.cfm?index=7122. Accessed March 3rd, 2009.

[30] Torrance, S. (2008): Ethics and Consciousness in Artificial Agents. *AI & Society*, 22 (4): 495-521.

[31] Wallach, W. & C. Allen (2009): *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

[32] Wallach, W., C. Allen, & I. Smit (2008): Machine Morality: Bottom-up and Top-down Approaches for Modelling Human Moral Faculties. *AI & Society*, 22: 565-582.

[29] Thank you to Verena Gottschling and Marcello Guarini for their help during the early stages of this research, and to two anonymous referees for their insightful feedback.

Faith in the Algorithm, Part 1: Beyond the Turing Test

Marko A. Rodriguez[1] and **Alberto Pepe**[2]

Abstract. Since the Turing test was first proposed by Alan Turing in 1950, the goal of artificial intelligence has been predicated on the ability for computers to imitate human intelligence. However, the majority of uses for the computer can be said to fall outside the domain of human abilities and it is exactly outside of this domain where computers have demonstrated their greatest contribution. Another definition for artificial intelligence is one that is not predicated on human mimicry, but instead, on human amplification, where the algorithms that are best at accomplishing this are deemed the most intelligent. This article surveys various systems that augment human and social intelligence.

The alleged short-cut to knowledge, which is faith, is only a short-circuit destroying the mind.

– Ayn Rand, "For the New Intellectual"

1 INTRODUCTION

The path towards artificial intelligence, in terms of mimicking human cognitive functionality, has been long, difficult, and at times it appears to have only made small steps. Bottom-up, state of the art vision systems have only accomplished modeling the functional capabilities of the V1, V2, and V4 regions of the visual cortex [33]. Popular, top-down knowledge representation and reasoning system are still primarily monotonic [26], are only beginning to incorporate common sense knowledge [28], and are predicated on logics that do not appear to model the true "rules" of human thought [38]. Moreover, these object recognition and knowledge representation and reasoning developments are but the fringe of a huge landscape of cognitive faculties that will be required to be simulated if human-type artificial intelligence is to ultimately be achieved in its fullest form. For example, other less developed agendas are object relation learning in neurally-plausible substrates [20], novel logic acquisition through experience [39], and associative mechanisms for merging the categorizations from different sensory modalities into a single language of thought [22]. Perhaps, by finding the lowest common denominator of the human neural system, it will be possible to simulate this behavior and expect for all other higher levels of intelligence to emerge through experience and learning. Modeling the processing capabilities of individual neurons has been the aim of the connectionist agenda for nearly three decades [32] and beyond various advances in classification, it appears that human type intelligence is still many more decades away. These statements serve not to criticize the researchers or their methods; rather, they are presented in order to acknowledge the level of difficulty involved in simulating human-type intelligence and the distances that need to be reached if this goal is to be achieved. Is it possible that the computer, and its underlying assumptions in logic, make it not that it is impossible to model human intelligence (assuming that such intelligence can be modeled on a Turing complete system), but instead blinds us as architects and engineers by biasing our approach? Moreover, is the Turing test [36] – the test for computer mimicry of a human – a red herring that is not a "natural" application of the computer's abilities? If so, what distinguishes human-type intelligence from that of a computers?

There are many designed tests that are used to quantify human intelligence. Interestingly, a human subject's scores in all of these tests have a positive correlation. Thus, regardless if a specialist is testing a subject's ability to manipulate objects in 3D space or the subject's fluency with language, success in one of these tests predicts success in another. This finding points to a single factor that can account for all human intelligence. This factor is known as the g-factor (or general intelligence factor) [35]. While being accepted as a single low-dimensional representation of human intelligence, it does not suffice to account for a true general theory of intelligence as those humans with savant syndrome (such as some autistics) demonstrate far-reaching intelligent abilities in one area, but not another; thus breaking the general applicability assumption of the g-factor [13]. Furthering this line of thought, the modern day computer can be seen as a savant in many respects. The computer demonstrates unmatched intelligence in very specific areas such as quickly translating languages during the process of compilation or in maintaining a lossless representation of a presented image in memory. Thus, in order to provide an account for intelligence beyond the g-factor (beyond the human as the golden measure), one can refer to the more general definition of intelligence – "doing the right thing at the right time" [29]. While this definition is not rigorous and provides no single quantifiable value defining a degree of intelligence, it articulates the general purpose of any intelligently behaving system. An understanding in terms of Darwinian natural selection elucidates that those systems that did the right thing at the right time continue to exist. Computers (more specifically their implemented algorithms) exist within this same natural selection framework where their evolution (through design) is pushing them to contribute to a previously unseen type of intelligence for a previously unseen type of environment. This intelligence is predicated on the integration of both the computer's and human's abilities. Moreover, and being the central thesis of this article, it is this type of intelligence that appears to be a more "natural" fit for the computer.

[1] Theoretical Division – Center for Non-Linear Studies, Los Alamos National Laboratory, email: marko@lanl.gov

[2] Center for Embedded Networked Sensing, University of California, Los Angeles, email: apepe@ucla.edu

2 HUMAN AND SOCIAL AUGMENTATION

Computers – the machines and their implemented algorithms – should not simply be interpreted as technological embodiments of solutions to specific problems. There is a larger relationship between the human, their problems and requirements, and designed algorithms and their executing hardware that are solving larger problems than either the human or the computer could solve alone; in other words, the computer is a contributing component within a larger intelligent system [18]. Sherry Turkle discusses the relationship between humans and computers as not just one in which the computer is a tool used to accomplish human tasks, but more of a component that works within the human's everyday life as a supporting entity [37]. From a "society of minds" perspective [27], the computer, as a cognitive component in human thinking, is very much a well functioning digital information processor much like the hippocampus is a well functioning neural memory device. In other words, the computer has found, not in any affective directed way, an information processing niche that further augments the human much like any other component of the human neural system [34]. To say whether the hippocampus is intelligent or not is to determine whether the results of its processing effect intelligent behavior; that is, does the human know where they are in physical space and do they encode episodic memories correctly. As an autonomous entity, the hippocampus, would appear, to the external human observer as not being intelligent at all. For one, in isolation, it simply becomes infected and its cells quickly die. However, within the larger schema of the human organism, its roll is of great significance to human intelligence; a few minutes interaction with the patient H.M. makes this point obvious [9]. Next, looking at the striate cortex demonstrates a relatively simple system [19] that implements a relatively simple algorithm (albeit on a massive scale) [33], but yet, when integrated within the nervous system as a whole, the contribution of the striate cortex to the overall intelligence of the human is immense. Without it, vision, and its associated functionalities, would not be possible; for instance, there would be no notion of a genius painter and the level of intelligence that such a connotation denotes. To this end, how many neural components are required before it is assumed that a human is intelligent? A review of the life and times of Helen Keller should demonstrate how vacuous this question is [24]. With an appeal to the Sorites paradox [8] and drawing, by analogy, from the late work of Ludwig Wittgenstein, what constitutes intelligence is one of "family resemblance" [42] and as such, a sharp definition is only grabbing at a vague notion. It is this argument that requires the loose definition of intelligence previously presented ("doing the right thing at the right time"). Any stricter definition would be riddled with exceptions.

Inevitably, this notion of intelligence needs to be situated within the context of the contemporary society, where the networked computer has permeated everyday life. This relationship, between the human and the computer in a technologically-driven society, unveils a natural symbiosis which is reminiscent of Hutchin's theory of distributed cognition [21] and to the notions of collective intelligence found in ant and termite populations [15]. Some of the tasks in which computers are employed in everyday life – from information access to social interaction – make this symbiosis evident. In many respects, traditional, "old fashioned" accounts of human intelligence (as evinced by the g-factor) refer to the emergent property of the coordinated activity of the individual's various brain regions. Introducing the computer into this system, a new type of intelligence emerges; an intelligence that, as argued, continues to maximize the general objective of doing the right thing at the right time – at both the individual and societal level.

The computer and its associated algorithms is a needed augmentation to the human individual given the number of options in the technologically-rich world and the difficulties in finding one's global optima within it. Moreover, society, in a collaborative fashion amongst its constituents and its supporting digital infrastructure, is making and will continue to make advances in the area of social intelligence, where an intelligent society is one that does the right thing at the right time. In this light, the question at hand is: what is the computer's contribution to intelligence? Or, in other words: in what ways have computers pushed humans and society into doing the right thing at the right time? In order to address this question, the following section explores the emergence of individual and social intelligence within the scope of the technological innovation that has most contributed to this type of augmentation in recent times: the World Wide Web.

3 EMERGENT WEB INTELLIGENCE

Since the dawn of the World Wide Web, information has been codified and distributed within a shared, universal medium that is accessible by human users world wide. The World Wide Web is unique for two reasons: distribution and standardization. In many respects, the first can not be accomplished without the latter. The Web's most eminent standard, the Uniform Resource Identifier (URI) has made it possible for the Web to serve as a network of information, from the document to the datum – a shared, global data structure [2]. This distributed data structure is amplifying the intelligence of the individual human and may provide a greater social intelligence. The remainder of this section will address the amplification of intelligence in the context of three general Web system: search engines (index and ranking), recommendation engines (personalized recommendations), and governance engines (collective decision making) [40].

3.1 Search Engines

The World Wide Web has emerged as a massive information repository in which humans contribute to and consume information from. This has not only provided humans a novel means of retrieving information, but also novel ways to publish and distribute information, thus leading to the increase in human information production. However, information increase inevitably brings about discoverability issues, as the necessity to locate and filter through desired information arises. To deal with this problem, algorithms have emerged to augment the individuals search capabilities. Interestingly, this augmentation is currently predicated on the contribution of many individuals within a stigmergetic environment [15].

The early Web maintained rudimentary indexes in the form of Web "yellow pages" that provided short descriptions of web pages. With the explosive growth of the Web, such directory services fell by the way side as no human operator (or operators) could keep up with the amount of information being published, nor could such rudimentary lists provide the end user the sophistication required to navigate the Web. By a nearly-Darwinian selection process, these early forms of indexes fell out of use because they were built around a conceptual framework that did not take advantage of the distributed representation of value inherent in every linking webpage made explicit by their authors. These rudimentary indexes of the early Web no longer function appropriately and as such, given the current requirements of the environment, are no longer able to do the right thing at the right time. As a remedy to this situation, a commercialized Web industry was

birthed and continues to thrive around solving the problem of search. Search engines index massive amounts of data that are gleaned from Web servers world wide. The development of the simple mechanism of ranking web pages by means of the their eigenvector component within the web citation graph has proved the most successful to date [7]. It is remarkable that this mechanism is entirely built around humans' decisions to link webpages; that is, the algorithm leverages human contributions and vice versa in a symbiotic manner. Even more remarkable is the fact that with the approximately 30 billion web pages in existence today, Web users can rest assured that, for the most part, their keyword search will provide the answer to their question within the first few results returned. This type of intelligence was not possible prior to the development of the Web, mainly because the problem of massive-scale indexing and ranking did not make itself apparent until the Web. However, this problem currently does exist and is being solved by the unification of the human's ability to, in a decentralized fashion, denote the value of web pages and in the computer's ability to calculate a global rank over these explicit expressions of value.

In this scenario, the Web plays the role of a digital Rolodex providing the human, nearly instantly, a reference to further information on nearly any topic imaginable [12]. Prior to the written document, information was passed from generation to generation in the form of large memorized stories and poems. In the contemporary technologically-rich world, this "algorithm" (cultural process) is no longer necessary. This is not to say that an individual can no longer memorize a long poem if they wished to, it is just that it is no longer required and as such, cognitive resources can be appropriated to other tasks as a new algorithm has emerged to handle this information indexing requirement. However, the Web is not a large story or poem; it follows no plot, no linear sequence, no poetic meter, no single language – the list of characters is beyond count and no single overarching writing style can be identified. For these reasons, it is posited that no currently existing neural component can memorize, index, and rank the entire Web, and thus as such, a specialized intelligence is required and has emerged.

3.2 Recommendation Engines

Large-scale human generated data sets have sowed the way for numerous algorithms. Such data sets includes the implicit valuation of resources that users leave on the web as they click from web page to web page or from purchased item to purchased item. No individual ever sees the entire Web and for the most part, for the life of the individual, they are confined to an ingrained path in a small subset of the greater Web as a whole. However, the aggregation of this clickstream information from all individuals provides a collectively generated representation of the inherent relationship between all items on the Web – from web pages to restaurants. This collective digital footprint provides not only novel ways to rank resources [5] but also, novel ways to recommend resources [6]. Other such human generated data are the numerous subjective ratings that individuals can provide on any topic imaginable – again, from webpages to restaurants. Finally, humans are also developing rich profiles of themselves that include not only identifiable facts such as one's curriculum vitae, but also the more qualitative aspects of someone's personality, tastes, and ever changing mood. There are many systems that take advantage of such data sets. A general application that is increasingly being used on such data sets is recommendation. A recommendation algorithm can be defined as any algorithm that provides users with resources (e.g. documents, books, music, movies, life partners, etc.) that are more likely than not to be correlated to the users' current requirements.

The popular collaborative filtering, recommendation algorithms of document and music services are able to utilize the previous click behavior of an individual, systematically compare it with the click behaviors of others, and from this comparison, recommend a set of resources that will be of interest to the user [17]. For many, the dependency on the librarian and the record shop owner has shifted to a dependence on this massive digital footprint and the algorithms that are able to utilize this footprint to the end user's advantage. The potential for the specialized intelligence of the computer to utilize complex mathematical approaches in clustering resources based on human behavior is something that no human can possibly accomplish.

An interesting phenomena to arise in recent years is the development and use of online dating services. In any large city there are too many individuals for any one human to sift through. Moreover, even if an individual had all the time in the world to meet everyone, the abilities of the individual may not be keen enough to predict, with any great accuracy, whether or not the one they are meeting will make an optimal mating partner. For this reason, dating services have emerged to handle, or rather attempt to handle, this common, pervasive problem. Ignoring broader social and cultural considerations for a moment, from a purely statistical perspective, the human's trial and error methods of sampling small portions of the population through friends or in social, physical environments (bars, restaurants, cafes, etc.) can not compete with the success rates of modern day matchmaking algorithms [1]. Note that matchmaking services are not something that is confined solely to the Web. Newspapers provide "personals" sections, but like the early "yellow pages" of the Web, they can not maintain rich human profiles, nor does manually browsing this information compare with the success of a matchmaking algorithm's recommendation. Again, for those activities for which a human simply does not have the skills to accomplish, the human relies on an external augmentation to fulfill the intelligence requirements of the problem at hand.

The recommendation services on the Web are following a common trend. They are all building more sophisticated models of the environment both in terms of the humans that utilize their services and in the resources that are indexed by these services. The World Wide Web infrastructure has provided the avenues for humans to collectively aggregate in a shared virtual space. Unfortunately, for the most part, the traffic data that is being generated as individuals move from site to site, the profiles that individuals repeatedly create at every online service, and the metadata about the resources that these services index are isolated within the data repositories of the services that utilize this information directly. Fortunately, recent developments in an open data model known as the "web of data" may change this by unifying the information contained in service repositories and exposing, within the shared, global URI address space, every minutia of data [4]. The end benefit of this shift in the perception of ownership and exposure of data will allow for a new generation of algorithms that take advantage of an even richer world model [25, 30]. Such models will include a seamless integration of the individual human's reading, listening, dating, working, etc. behavior as well as the descriptions of books, songs, movies, people, jobs, etc. At this point, to the algorithms that leverage such data, a human is no longer just a consumer of a particular type of literature or a connoisseur of a particular style of film, but rather, a complex entity that can be subtly oriented, through recommendation, in a direction that ensures that they are experiencing that aspect of the world that is most fitting to who they are

at the moment that they are that.

At the extreme of this line of thought, if enough information is gathered and a rich enough world model is generated, then it may be possible to design algorithms that are more fit to determine the life course of an individual human than what the individual, their family, or their community can do for them (with appropriate feedback from the world to the model [14], which may include the perspectives of the individual, their family, and their community). This view suggests that it may be best to rely on a large-scale world model (and algorithms that can efficiently process it) when making decisions about one's path in life. Such algorithms can take into account the multitude of relations between humans and resources, and improvise a well "thought out" plan of action that ensures that the individuals, to the best of the system's ability, live a life that is filled with optimal experiences – of experiences where they did the right thing at the right time; a life in which the others they met, the restaurants they frequented, the books they read, the classes they attended, and so forth led to experiences that were completely fulfilling to the human. These optimal experiences represent the perfect balance between the psychological states of anxiety and boredom and as such, would increase the individuals' attentiveness and involvement in such activities – similar to the mental state that is colloquially known as "flow" [10].

A large-scale world model has the potential to integrate the collective zeitgeist of a society, the socio-demographic and geographic layouts of cities, the location of its inhabitants, their personal characteristics, their resources and relations. Amazingly, such data currently exist in one form or another, to varying degrees of accuracy, completeness, and levels of access. Further making this information publicly available and integrated would allow for algorithms (under Darwinian selection processes) to evolve, over iterations of development and insight, that were fit to determine the individuals' global optima. At this level of life optimization, it could be argued that a maximally intelligent human has emerged – a life (as subjectively interpreted by the individual) that was filled with moments where the right thing was always being done at the right time.

3.3 Governance Engines

In many ways, aiding the human in finding global optima is the purpose of a society (within the constraints of taking into account the optima of others). From high-level governmental decisions to the local cultural rules that determine the way in which humans interact in their environment, the goal of a (benevolent) society is to ensure a life in "the pursuit of happiness" [23]. The question is then: what are the limits of happiness and well-being that can be achieved by the current societal structures alone? And also: are there more efficient and accurate algorithms that can be utilized to ensure the greatest benefit to human life? Recommendation systems are a step in the direction towards the use of computers to provide the human the right resource at the right time, regardless of what form that resource may take. However, within the grander scheme of society as a whole, the nascent fields of e-governance and computational social choice theory are only beginning to tangentially touch upon the idea that a networked computer infrastructure could be used to foster a new structure for government.

Reflecting on modern voting mechanisms (specifically those within the United States), we find a system that is fragile, inaccurate, and expensive to maintain. Due in part to the outdated infrastructure that citizens use to communicate with their governing body, citizen participation in government decision making is limited. However, these days, with the level of eduction that citizens have, the amount of information that citizens can become aware of, and the sophistication of modern network technologies, is it possible that current government decisions are limited in that they are not leveraging the full potential of an enlightened population (or subset thereof)? By making use of both a large-scale and knowledgeable decision making constituent, it is theoretically possible that all decisions, made by the decision making constituency, are optimal. This statement was validated (under certain assumptions) in 1776 by Marquis de Condorcet's famous Condorcet jury theorem [11].

With the social networks that are being made explicit on the Web today, and with open standard movements that ensure that this information can be shared across services, it is possible to leverage a relatively simple vote distribution mechanism to remove the representative layers of governance and promote full citizen participation in all the decision making affairs of a society. This mechanism, known as dynamically distributed democracy, ensures that any actively participating subset of a population simulates the decision making behavior of the whole [31]. Thus, a simulated, large-scale decision making body can be leveraged in all decisions. A large decision making body is the first requirement of the Condorcet jury theorem. Next, Robin Hanson articulates a vision of government where any individual can participate through a decision system known as a prediction market [16]. The purpose of a prediction market is to provide accurate predications of objectively determinable states of the world (current or into the future) and its application to governance is noted in the popular phrase "vote on values, but bet on beliefs." In this form, the self-selecting, monetary mechanisms that determine whether someone participates is based on their degree of knowledge of the problem space. Those that are not knowledgeable, either do not participate or loose money in the process of participating; thus, hampering the individual from participating in matters outside the scope of their abilities into the future. The accuracy of such systems are astounding and have popular uses in election predictions and a short lived run in terrorist predictions (only to be dismantled by the U.S. government because it was considered too morose for market traders to monetarily benefit on the accurate prediction of the death of others). A knowledgeable decision making body is the second requirement of the Condorcet jury theorem and, much like commodity markets, prediction market systems select for knowledgeable individuals.

These ideas stress the importance of reflecting on the medium by which society organizes itself, generates its laws, and implements methods in how it will utilize resources most effectively. Like the "yellow pages" of the early Web, it may not be optimal to leave such pressing matters to an operator (or operators). That statement is not a critique of the leaders and doctrines of nations, but instead is a comment on the complexity of the world and the necessity for a new type of intelligence; moreover, it is posed as an appeal to rethink government and its role within contemporary networked society [3]. A distributed value/belief system and algorithmic aggregation mechanism may prove to be the better problem-solving mechanism for societal issues into the future. It is in this area that computers, with their savant-like abilities, may contribute to social intelligence, where the unification of the intelligence augmentation gained by the individual human and the society coalesce into a type of intelligence that is novel (beyond human mimicry) and above all beneficial.

4 CONCLUSION

Humans perceive their world through their sense modalities, create stable representations of the consistent patterns in the world, and uti-

lize those representations to further act and survive to the best of their abilities. Their internal, subjective world is an endless stream of thoughts – a complex, information-rich map of the external world [41]. Manifestations of intelligence – "doing the right thing at the right time" – inherently depend upon an individual's internal representation of the external world. By analogy to the field of computer science, this internal map of the world can be regarded as the data structure upon which reasoning mechanisms (i.e. algorithms) function. From an objective perspective, the human mind can only maintain so rich a data structure, process only so many aspects of it, and simulate only so many potential future paths for the individual to choose from. The complexity of the human's mental calculation grows when considering that many other such simulations are occurring in the minds of their fellow men and women and like a general-purpose processor, to simulate a machine within a machine reduces the resources available to the original machine to execute other processes. For these reasons, the human is not a perfectly intelligent creature always doing the right thing at the right time.

As discussed, with the externalization of the human's internal world through the explicit expression of themselves, their relation to others, and the resources with which they rely upon, other processes can utilize this explicit model to aid the human in the process of life and thus, the process of thought. The World Wide Web and the algorithms implemented upon it function like an auxiliary mind, exposed to more information than could be possibly processed by its neural counterpart. While the core specification of these algorithms may be understood, even thoroughly by their designers, ultimately what machines compute are based on such a large-scale model of the world, that to assimilate its results into one's choices are ultimately based on faith – much like the faith one has in the validity of their episodic memories and their current location in space as provided to them by their hippocampus.

REFERENCES

[1] Aaron Ben-Ze'ev, *Love Online: Emotions on the Internet*, Cambridge University Press, 2004.

[2] Tim Berners-Lee and James A. Hendler, 'Publishing on the Semantic Web', *Nature*, **410**(6832), 1023–1024, (April 2001).

[3] Colin Bird, 'The possibility of self-government', *The American Political Science Review*, **94**(3), 563–577, (2000).

[4] Christian Bizer, Tom Heath, Kingsley Idehen, and Tim Berners-Lee, 'Linked data on the web', in *Proceedings of the International World Wide Web Conference WWW08*, Workshop on Link Data (LDOW2008), Beijing, China, (April 2008).

[5] Johan Bollen, Herbert Van de Sompel, and Marko A. Rodriguez, 'Towards usage-based impact metrics: first results from the MESUR project.', in *Proceedings of the Joint Conference on Digital Libraries*, pp. 231–240, New York, NY, (2008). ACM Press.

[6] Johan Bollen, Michael L. Nelson, Gary Geisler, and Raquel Araujo, 'Usage derived recommendations for a video digital library', *Journal of Network and Computer Applications*, **30**(3), 1059–1083, (2007).

[7] Sergey Brin and Lawrence Page, 'The anatomy of a large-scale hypertextual web search engine', *Computer Networks and ISDN Systems*, **30**(1–7), 107–117, (1998).

[8] James Cargile, 'The Sorites Paradox', *British Journal for the Philosophy of Science*, **20**(3), 193–202, (1969).

[9] Neal J. Cohen, *Memory, Amnesia, and the Hippocampal System*, MIT Press, September 1995.

[10] Mihály Csíkszentmihályi, *Flow: The Psychology of Optimal Experience*, Harper and Row, New York, NY, 1990.

[11] Marquis de Condorcet. Essai sur l'application de l'analyse á la probabilité des décisions rendues á la pluralité des voix, 1785.

[12] Douglas C. Engelbart, *Computer-supported cooperative work: a book of readings*, chapter A conceptual framework for the augmentation of man's intellect, 35–65, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[13] Howard Gardner, *The origins and development of high ability*, chapter The relationship between early giftedness and later achievement, John Wiley and Sons, 1993.

[14] Vadas Gintautas and Alfred W. Hübler, 'Experimental evidence for mixed reality states in an interreality system', *Physical Review E*, **75**, 057201, (2007).

[15] P. Grasse, 'La reconstruction du nid et les coordinations interindividuelles chez bellicositermes natalis et cubitermes sp. la theorie de la stigmergie', *Insectes Sociaux*, **6**, 41–83, (1959).

[16] Robin Hanson, 'Shall we vote on values, but bet on beliefs?', *Journal of Political Philosophy*, (in press).

[17] Johnathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl, 'Evaluating collaborative filtering recommender systems', *ACM Transactions on Information Systems*, **22**(1), 5–53, (2004).

[18] Francis Heylighen, 'The global superorganism: an evolutionary-cybernetic model of the emerging network society', *Social Evolution and History*, **6**(1), 58–119, (2007).

[19] D. H. Hubel and T. N. Wiesel, 'Receptive fields and functional architecture of monkey striate cortex.', *Journal of Physiology*, **195**(1), 215–243, (March 1968).

[20] J.E. Hummel and K.J. Holyoak, 'A symbolic-connectionist theory of relational inference and generalization', *Psychological Review*, **110**(2), 220–264, (2003).

[21] Edwin Hutchins, *Cognition in the Wild*, MIT Press, September 1995.

[22] Ray S. Jackendoff, *Languages of the Mind*, MIT Press, 1992.

[23] Thomas Jefferson. Declaration of independence, 1776.

[24] Helen Keller, *The Story of My Life*, Doubleday, Page and Company, New York, NY, 1905.

[25] Lawrence Lessig, *Free Culture: The Nature and Future of Creativity*, CreateSpace, Paramount, CA, 2008.

[26] Deborah L. McGuinness and Frank van Harmelen. OWL web ontology language overview, February 2004.

[27] Marvin Minsky, *The Society of Mind*, Simon and Schuster, March 1988.

[28] Erik T. Mueller, *Commonsense Reasoning*, Morgan Kaufmann, January 2006.

[29] Tony J. Prescott, Joanna J. Bryson, and Anil K. Seth, 'Modelling natural action selection', *Philosophical Transactions of the Royal Society B: Biological Sciences*, **362**(1485), 1521–1529, (September 2007).

[30] Marko A. Rodriguez, 'A distributed process infrastructure for a distributed data structure', *Semantic Web and Information Systems Bulletin*, (2008).

[31] Marko A. Rodriguez and Daniel J. Steinbock, 'A social network for societal-scale decision-making systems', in *Proceedingss of the North American Association for Computational Social and Organizational Science Conference*, Pittsburgh, PA, USA, (2004).

[32] David E. Rumelhart and James L. McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press, July 1993.

[33] Thomas Serre, Aude Oliva, and Tomaso Poggio, 'A feedforward architecture accounts for rapid categorization', *Proceedings of the National Academy of Science*, **104**(15), 6424–6429, (April 2007).

[34] Peter Skagestad, 'Thinking with machines: Intelligence augmentation, evolutionary epistemology, and semiotic', *Journal of Social and Evolutionary Systems*, **16**(2), 157–180, (1993).

[35] Charles Spearman, 'General intelligence objectively determined and measured', *American Journal of Psychology*, **15**, 201–293, (1904).

[36] Alan M. Turing, 'Computing machinery and intelligence', *Mind*, **58**(236), 433–460, (1950).

[37] Sherry Turkle, *The Second Self: Computers and the Human Spirit*, MIT Press, 1984.

[38] Pei Wang, 'Cognitive logic versus mathematical logic', in *Proceedings of the Third International Seminar on Logic and Cognition*, (May 2004).

[39] Pei Wang, *Rigid Flexibility*, Springer, 2006.

[40] Jennifer H. Watkins and Marko A. Rodriguez, *Evolution of the Web in Artificial Intelligence Environments*, chapter A Survey of Web-Based Collective Decision Making Systems, 245–279, Studies in Computational Intelligence, Springer-Verlag, Berlin, DE, 2008.

[41] Ludwig Wittgenstein, *Tractatus Logico-Philosophicus*, Routledge, September 1922.

[42] Ludwig Wittgenstein, *Philosophical Investigations*, Blackwell Publishers, April 1973.