Computing & Philosophy

AISB 2008

AISB 2008 Proceedings Volume 12

GLoRiClass

AISB ABERDEEN - 2008

UNIVERSITY OF ABERDEEN

AISB 2008 Convention
Communication, Interaction and Social Intelligence

1st-4th April 2008

University of Aberdeen

Volume 12 :

Proceedings of the

AISB 2008 Symposium on Computing and Philosophy

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

http://www.aisb.org.uk/convention/aisb08/

ISBN 1 902956 71 0

Contents

The AISB'08 Convention ii
Frank Guerin & Wamberto Vasconcelos

Symposium Preface iii
Mark Bishop

What would a Wittgensteinian computational linguistics be like?? 1
Yorick Wilks

Cognition without content 7
Paul Schweizer

Foundations of a Philosophy of Collective Intelligence 12
Harry Halpin

Constructivism in AI: Prospects, Progress and Challenges 20
Frank Guerin

Social Robotics and the person problem 28
Stephen J. Cowley

The Antiquarian Librarian & the Pedantic Semantic Web Programmer: Trust, logic, knowledge and inference 35
Cate Dowd

Could a Created Being ever be Creative? Some Philosophical Remarks on Creativity and AI Development 43
Y. J. Erden

A Modelling Framework for Functional Imagination 51
Hugo Gravato Marques, Owen Holland & Richard Newcombe

The Plaited Structure of Time in Information Technology 59
Ganascia Jean-Gabriel

Substitution for Fraenkel-Mostowski foundations 65
Murdoch J. Gabbay & Michael J. Gabbay

The AISB'08 Convention: Communication, Interaction and Social Intelligence

As the field of Artificial Intelligence matures, AI systems begin to take their place in human society as our helpers. Thus it becomes essential for AI systems to have sophisticated social abilities, to communicate and interact. Some systems support us in our activities, while others take on tasks on our behalf. For those systems directly supporting human activities, advances in human-computer interaction become crucial. The bottleneck in such systems is often not the ability to find and process information; the bottleneck is often the inability to have natural (human) communication between computer and user. Clearly such AI research can benefit greatly from interaction with other disciplines such as linguistics and psychology. For those systems to which we delegate tasks: they become our electronic counterparts, or agents, and they need to communicate with the delegates of other humans (or organisations) to complete their tasks. Thus research on the social abilities of agents becomes central, and to this end multi-agent systems have had to borrow concepts from human societies. This interdisciplinary work borrows results from areas such as sociology and legal systems. An exciting recent development is the use of AI techniques to support and shed new light on interactions in human social networks, thus supporting effective collaboration in human societies. The research then has come full circle: techniques which were inspired by human abilities, with the original aim of enhancing AI, are now being applied to enhance those human abilities themselves. All of this underscores the importance of communication, interaction and social intelligence in current Artificial Intelligence and Cognitive Science research.

In addition to providing a home for state-of-the-art research in specialist areas, the convention also aimed to provide a fertile ground for new collaborations to be forged between complementary areas. Furthermore the 2008 Convention encouraged contributions that were not directly related to the theme, notable examples being the symposia on "Swarm Intelligence" and "Computing and Philosophy".

The invited speakers were chosen to fit with the major themes being represented in the symposia, and also to give a cross-disciplinary flavour to the event; thus speakers with Cognitive Science interests were chosen, rather than those with purely Computer Science interests. Prof. Jon Oberlander represented the themes of affective language, and multimodal communication; Prof. Rosaria Conte represented the themes of social interaction in agent systems, including behaviour regulation and emergence; Prof. Justine Cassell represented the themes of multimodal communication and embodied agents; Prof. Luciano Floridi represented the philosophical themes, in particular the impact of society. In addition there were many renowned international speakers invited to the individual symposia and workshops. Finally the public lecture was chosen to fit the broad theme of the convention – addressing the challenges of developing AI systems that could take their place in human society (Prof. Aaron Sloman) and the possible implications for humanity (Prof. Luciano Floridi).

The organisers would like to thank the University of Aberdeen for supporting the event. Special thanks are also due to the volunteers from Aberdeen University who did substantial additional local organising: Graeme Ritchie, Judith Masthoff, Joey Lam, and the student volunteers. Our sincerest thanks also go out to the symposium chairs and committees, without whose hard work and careful cooperation there could have been no Convention. Finally, and by no means least, we would like to thank the authors of the contributed papers – we sincerely hope they get value from the event.

Frank Guerin & Wamberto Vasconcelos

The AISB'08 Symposium on Computing and Philosophy

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 modeling of philosophical assumptions. To this end the group of topics selected for discussion in the symposium include: Natural Language Processing - examining philosophical assumptions and the connections that can be surmised between those and views about language traditionally associated with Wittgenstein; Cognition and content; constructivism; social robotics and collective intelligence; the semantic web; computational creativity; computational imagination; the structure of time and logical substitution.

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

Mark Bishop
Goldsmiths, University of London, UK.

Programme Chair:
Mark Bishop (Goldsmiths, University of London, UK)

Organising Committee:
Peter Baumann (Aberdeen University, UK)
Mark Bishop (Goldsmiths, University of London, UK)
Luciano Floridi (University of Hertfordshire & St. Cross College
Steve Torrance (Middlesex University & University of Sussex, UK)

Programme Committee:
Peter Baumann (Aberdeen University, UK)
Ron Chrisley (University of Sussex, UK)
Luciano Floridi (University of Hertfordshire & St. Cross College Oxford University, UK)
John Preston (University of Reading, UK)
Murray Shanahan (Imperial College, UK)
Colin Schmidt (Universit du Maine, France)
Keith Stenning (The University of Edinburgh, UK)
Susan Stuart (The University of Glasgow, UK)
Steve Torrance (Middlesex University & University of Sussex, UK)
Michael Wheeler (University of Stirling, UK)

What would a Wittgensteinian computational linguistics be like?

Yorick Wilks[1]

Abstract. The paper tries to relate Wittgenstein's later writings about language with the history and content of Artificial Intelligence (AI), and in particular, its sub-area normally called Computational Linguistics, or Natural Language Processing. It argues that the shift, since 1990, from rule-driven approaches to computational language and logic, associated with traditional AI and the linguistics of Chomsky, to more statistical models of language have made those connections more plausible, in particular because there is good reason to think the latter is a better model of use than the former. What statistical language models are not, of course, are immediately plausible models of meaning. Moreover, a statistical model seeking a model of a whole language, one can now look at the World Wide Web (WWW) as an encapsulation of the usage of a whole a language, open to computational exploration, and of a kind never before available. I describe a recent empirical effort to give sense to the notion of a model of a whole language derived from the web, but whose disadvantage is that that model could never be available to a language user because of the sheer size of the WWW. The problematic issue in such an analogy (Wittgenstein and NLP) is how one can go beyond the anti-rule aspect of both to some view of how concepts can even appear to exist, whatever their true status.

"A main source of our failure to understand is that we do not command a clear view of the use of our words – our grammar is lacking in this sort of perspicuity. A perspicuous representation produces just that understanding which consists in "seeing connexions". Hence the importance of finding and inventing intermediate cases. The concept of a perspicuous representation is of fundamental significance for us. It earmarks the form of account we give, the way we look at things." Wittgenstein: Philosophical Investigations $122. (My emphasis)

1 INTRODUCTION

Seeking out its intellectual roots or scholarly ancestors is not an activity popular or respected in the technology called Natural Language Processing (NLP, alias Computational Linguistics [1]). Many of its researchers have some vague notion that logical predicate representation, now almost a form of shorthand in NLP, owes a lot to Frege and Russell, but few know or care that, long before Chomsky ([2], if we agree to allow him by courtesy into the history of NLP) Carnap, Chomsky's teacher, set up in the 1930s what he called The Logical Syntax of Language ([3]) with formation and transformation rules whose function was to separate meaningful from meaningless expressions by means of rules. Carnap's driving role behind all that has been utterly forgotten and Chomsky's own work has now simply filled in all the intellectual space in formal linguistics.

Another contemporary of Carnap, also now largely lost to view, is Wittgenstein, whose long campaign against simple-minded notions of linguistic rules was largely provoked by Carnap. He predated Chomsky and NLP, of curse, although his influence lived on as a source of Anglo-Saxon linguistic philosophy for many decades, whose practitioners mostly had little time or patience for what they saw as Chomsky's simplicities and certainties.

An attempt to connect Wittgenstein to linguistics thirty years ago was Brown's "Wittgensteinian Linguistics" [4], but his main concern was to contrast Wittgenstein with Chomsky's views, which were more central to language studies then than they are now. Brown noted that Wittgenstein had much in common with Chomsky's anthropological predecessors, from whom he separated himself so clearly with his rule-driven, Carnap-inspired linguistics. Malinowski's observation ([5]:287ff) that language is "a mode of action, rather than a counter-sign of thought" is a sentiment that Wittgenstein could have expressed, and the latter's notion of communities of use who share assumptions and language forms, however bizarre, is not far from anthropological views (often associated with Whorf and Sapir) on the language and belief systems valid in their own terms. Quine [6] later took up the same scenario, that of remote languages, unknown to the observer, and the non-veridical nature of any communication based on translation or supposed meaning equivalence: how could we ever know definitively, he asked, what "Gavagai" meant simply from the utterances (and pointings) we observed?

Wittgenstein seemed less sceptical about translation than Quine; perhaps living in two languages and cultures, as he did, made it seem more natural to him: classic sentiments like "the limits of my language mean the limits of my world" ([7]) do not imply that one cannot be in two or more such worlds. He listed (PI [8] pp.11-12) translation as among normal human activities, and he seemed sceptical about the nature and function of none of his list. It also seems clear that Wittgenstein did believe in some conceptual world over and above surface use, but the problem is knowing what that was, and how it was grounded within usage. In his early work, what he called forms of facts [7] were separate from language and identified with "pictures of fact" and it is not clear that he ever rejected the explanatory power of diagrams and pictures: he continued to use them, even though he was unsure how they "worked" (cf. The problem of knowing why the arrow so obviously points the way it does [8] PI: (129). Pictures and drawings remained important to Wittgenstein because they expressed intention in a way that natural objects in the world do not.

[1] Oxford Internet Institute, University of Oxford.
yorick.wilks@oii.ox.ac.uk

In spite of many things he says that appear to be classic behaviourism –e.g. the apparent denial of the possibility of a private language – Wittgenstein was not an empiricist in the sense that Chomsky intended by that word, as is someone like Sampson [9] who insists that we have no evidence that anything more is innate in humans than a learning mechanism. Wittgenstein could never have written "It is conceivable....that all the processes of understanding, obeying, etc. should have happened without the person ever having been taught the language" (PI $12) had that been his position. Chomsky himself seems to have no understanding whatever of Wittgenstein's overall position, given remarks like: (Chomsky [10]:p60) "[For Wittgenstein] meanings of words must not only be learned, but also taught (the only means being drill, explanation, or the supplying of rules….." . Chomsky has no feeling at all for Wittgenstein's investigation of how we could know that someone was following [a linguistic] rule, and for the simple reason that Chomsky always claims to know that we are following rules, and when, and to see no problem about a statement that a rule is being followed by a speaker.

These arguments, that effectively separate Wittgenstein in every way from the Chomskyan enterprise, can be found in Brown's work, but one must add here that Chomsky and classic Artificial Intelligence (AI, e.g. [11])—with its emphasis on the role of logic as a "mental representation" – -are not very different positions when contrasted to Wittgenstein. However, our focus here will be to contrast and compare Wittgenstein with developments specifically in NLP and computational linguistics, which has become more central within linguistics as a whole, as Chomsky's influence has declined, and not with the rule-driven paradigm (of Chomsky and in its different way, classical Artificial Intelligence) but with the more statistical paradigm that has replaced it.

Since Brown, one can hear new echoes in NLP of Wittgenstein's influence, as when Veronis called recently for looking "not for the meaning but the use" [12], thus reviving one of the best known Wittgensteinian slogans. One could hear it, too, in Sinclair's call to let a corpus "speak to one" [13], without the use of analytical devices and in Hanks' claim [14] that a dictionary could be written consisting only of use citations. This last may well be false, for it is hard to see the function of a dictionary that did not explain, but it does contain the authentic Wittgensteinian demand to look at language data, even if not in the way a linguist would mean who gave the same exhortation (i.e. to form a generalization from it, in the linguist's case).

Wittgenstein, of course, knew nothing of computers in the modern sense, although he trained as an engineer. All I can do in this brief paper is to more or less assume his views on language known to the reader, and to note what movements in modern NLP those are closer to and farther from, and why his arguments and insights should still be taken account of by those concerned to process language by machine. This paper will not be about scholarly claims of direct influence, for there are probably few to be found. Margaret Masterman [15] and perhaps the present author are two of the very few NLP researchers who acknowledged his influence and referred to him often. One thinks here, too, of Graeme Hirst's immortal and not wholly serious: "Most Artificial Intelligence programs are in Wittgenstein and only the degree of implementation varies", which only serves to show how much remedial work there is to do.

2 THE WORLD WIDE WEB AS A CORPUS OF USE

Wittgenstein's appeal to look for the use rather than the meaning is not, on its face, a clear injunction: elsewhere he writes of giving meanings by means of explanations (Blue Book [16] p27) and one may reasonably infer that the meanings NOT to look for are pointings at objects, and that when meanings are to be given they are in terms of more words, paraphrases (and not, he makes clear elsewhere, definitions) rather than an artificial coded language for meaning expression, such as that traditionally offered by logic, and later by linguistics and AI.

All this suggests an approach to actual language use more sympathetic than that usually associated with philosophers, and that was indeed the movement he created. Later, Quine, who made many of the same assumptions as Wittgenstein, explicitly linked looking at language use with the methods of structural (i.e. pre-Chomskyan or anthropological) linguistics, seeking data in languages not understood by the researcher, and drew a range of conclusions [6] very close to those of Wittgenstein, in particular that it was not mere language data that would do the trick but data in a language that was understood, by whatever process.

This also shows how wary one must be of trying, as Brown did, to place Wittgenstein somehow closer to the anthropological-empirical tradition than to Chomsky. It is true that Wittgenstein had something in common with the earlier writers, as Brown noted, but his emphasis on seeing language "from the inside', as something already understood and distinctively human, rather than as an object for scientific observation, brings him closer to Chomsky's emphasis on the native speaker and intuition. The truth is that, while Chomsky was a committed anti-behaviourist, Wittgenstein maintained an ambiguous position, one which declined to give the speaker veridicality on what he meant, so that he could not be wrong, a certainty Wittgenstein considered vacuous.

Among those who traditionally drew the attention of NLP researchers to data in large quantities were lexicographers, of linguistic or computational bent, as the remarks of Sinclair and Hanks above show. Since the return of machine learning and statistical methods to NLP, applied to large corpus data bases since the early 1990s, and following their proven success in speech recognition, NLP has taken large collections of text seriously as its databases. Recently, Kilgarriff and Grefenstette [17] based a journal issue on the notion of "web as corpus": the use of the whole web in a given language as a corpus for NLP and, given Grefenstette's estimates [18], it is now clear the total of pages in English is up to forty times the number indexed by Google (currently more than 10 billion).

A corpus of that size is of course a data base of use/usage, one far greater than any human could encounter in a lifetime, and it is not structured in the way any human would encounter language, e.g. as dialogue, rather than prose, and graded appropriately for age on encountering it. But, of course, that is just a search problem, too, for there must be, in those 300 billion pages of English, a great deal of dialogue and child language at all levels. We must give up any idea that such a vast corpus could be a cognitive model of any kind: it would take a reader, reading constantly, at least 60,000 years to train on the current English web corpus. One can compare this with Roger Moore's observation that [19] if a baby learned to speak using the best

models of speech acquisition currently available, it would take 100 years to learn to talk.

The question we can now ask is, does that access to the whole web as a corpus by NLP research bring us closer to an ability to compute over usage in a language as a whole, to language surveyed in its full variety, rather than the examples an individual might think up, or generate from rules, or whatever? The odd answer seems to be that, although a web corpus, even now, only fifteen years after its inception, is so vast in human-life (i.e. of reading) terms, it is still no kind of full survey of language possibilities and never can be. And the reason for that lies not any kind of Chomskyan notion of novelty to do with the infinite number of sentences that can be generated from a finite base of rules.

For there is no finite base in any straightforward sense: as far as words (what some call unigrams) are concerned, it is clear they will continue to occur at a steady rate no matter how large the corpus [20]. This fact also holds for all forms of combinations of words. These are only examples of what is known as "data sparseness", and maybe no more than a statistical/combinatoric updating of Chomsky's point: as Jelinek has put it from a statistical point of view: "language is as system of rare events". But is it vital to emphasise (since this whole discussion will have to be brought back to the notion of rules in due course) how wrong that finite base assumption of Chomsky's was. Krotov induced all possible phrase-structure rules explicitly from the large passed corpus called the Penn Tree Bank (PTB) and plotted them against the length of the (part of) the corpus that gave rise to them. What was clear and astonishing was that at the end of the process –i.e. training on the whole of the PTB – the number of rules found (over 18K) was still rising linearly with the length of the corpus! It is quite unclear that there is any empirical justification for the idea of a finite syntactic base, at least for English, for that would require that that graph flattened at some point. This suggests that any rule base will continue to grow indefinitely with a new corpus, just as the (unigram) vocabulary does. There is no reason to think this tendency will change with much longer corpora; given that fact, assuming it is one, it is one hard to grasp within the history of modern formal linguistics. Chomsky took it simply as an article of faith that there must be a finite set of rules underlying a language, if only they could be written or found [21] suggests this is simply not so.

We are approaching a paradox here: there is an opposition, clear in Wittgenstein, to the notion of boundedness of a language implied by the rule-driven approach to a natural language, one found in Carnap, and which continued in Chomsky's work. Wittgenstein wanted to question both that we could be said to be using any such rules and that any set of them could bound a language and determine well-formedness. Goedel's results on undecidability in mathematics [22] must have seemed to him analogues from that world, and this is explicit in the Remarks on the Foundations of Mathematics [23] see also [24]).

However, just as it may be the case that the rule set for a language, like its sentence set, is not finite at all, so it may be the case that a corpus, for a language itself cannot be bounded, no matter how large it grows; or, rather, there is no corpus that captures the whole language, and so usage/use itself it not something finite that can be appealed to. One could, presumably, restrict oneself to all the sentences of English up to, say, 15 words long and bound that by permutations, but the problem remains that the word set itself is shifting all the time: e.g. more than 900 words a year are being added to non-scientific English (The Times, 9/10/03).

Can Wittgenstein's appeal to use be related to the fact that NLP over the whole web now surveys enormously more use than it did? It is clear that there can now be real experiments that appeal to use in a very satisfying way: Grefenstette, for example, (18) has described a novel algorithm for machine translation – following an earlier suggestion due to Dagan – in which a (two word) Spanish bigram XY is translated into, say, English by taking the n senses of a Spanish word X in a Spanish-English bilingual dictionary, and making a Cartesian product with the m senses of Spanish word Y, and then seeking the n x m resulting English bigrams in an English corpus and ranking them by frequency of occurrence. One may be confident that the most frequent one is always the correct translation.

This algorithm is in fact quite hard to explain and justify a priori: it feels exactly like "Asking the audience" in the popular quiz show "Who wants to be a Millionaire?" where, again, the most frequent answer from the audience is usually, but not always, correct, a phenomenon very close to what some would call the Google-view-of-truth, or what is now referred to as the "Wisdom of Crowds" But whatever is the case about that, there is no doubt this algorithm is precisely an appeal to use rather than meaning and a model for the future deployment of the web-as-corpus to solve linguistic problems.

3 BACK TO THE PRESENT STATE OF CL/NLP

Let us turn back now to the state of computational linguistics and NLP by computer. One could generalize very rapidly as follows: in the 1970's, there arose movements such as Schank's conceptual dependency or preference semantics (Wilks, [26]) which could be described as attempting to map a "deep grammar' of concepts and what I would call the preferential relations between concepts. This theme was closely allied with various forms of Fillmore's [27] case grammar in linguistics, and his more recent work [28] could certainly be described as a continuing search for local, but deep, grammatical relations – based on systematic substitution relations in semi-fixed phrases in English – outside the concerns of the main thrust of work in computational syntax, which is little concerned with words themselves or local effects in language. Fillmore's hand-coded lexicography just mentioned has been a survivor, but virtually all other attempts at conceptual mapping have been overtaken by one of the two separate movements to introduce empiricism into CL and NLP: the connectionist movement of the early 1980s, and the statistical corpus movement, driven by Jelinek's successes in speech and then translation in the late 1980s. [29] The first was not a success but the second is still continuing: a classic of the first movement would be Waltz and Pollack's [30] neural networks showing how concepts attracted and repelled each other in terms of contexts supplied to the network, from corpora or from dialogue. The work was exciting but such networks were never able to process more than tiny fragments of language. There were more radical (or "localist") connectionists. such as ([31] who went further and declined to start from explicit language symbols at all, in an attempt to show how symbols could have been reached from simpler associationist algorithms that built, rather than assumed, the symbols we use. If this had

been done it might have broken through the impasse that the title of this paper suggests, namely how can one have a theory of language which does not build in from the very start all that one seeks to explain, as intuition-based theories in linguistics, logic and AI always seem to. Connectionist theories could never give a clear account of the theory-free "simples" from which to begin, and in any case they also failed to "scale up" to any reasonable sample of language use, or to confirm any strong claims about human cognition of language.

The second movement, that followed connectionism, the one we are still within, at the time of writing, was statistical associationism, driven by Jelinek with his translation work derived from trigram models of speech [29], and which had some success and undoubtedly used language on a very large scale indeed, too large as we noted earlier, to be cognitively plausible for human beings. This movement has been committed to an "empiricism of use" but can such approaches ever build back to reconstruct concepts empirically? This movement, as we noted earlier, shares many assumptions with the technology of Information Retrieval (IR) (see e.g. [32]): a view that language consists only of words without the meta-codings that concepts and linguistic features" claim to provide, and that all the decorations and annotations that intuitive theories add are unexplained and unacceptable as explanatory theory. IR, it must always be remembered, underlies the successful search theories that have given us the World Wide Web search tools.

After his surprisingly successful machine translation project at IBM, done only by statistics, Jelinek became disillusioned with his first set of statistical functions and came to the view that language data is too sparse to allow the derivation of what he called a full trigram model of a language, which is to say, one that would have to be derived from a corpus so large that one could expect to have seen, when training on it, every trigram one could find in any text being tested subsequently—every possible sequence of three words the language allows (three here being an arbitrary number, cut off at a level where computation is unfeasible for larger numbers).

If I present this paper at the AISB, I will at this point briefly describe some recent experimental work with colleagues at Sheffield that suggests that Jelinek may have been too pessimistic, and that a full trigram model might now be within reach, using a device called a "skipgram". These are "trigrams with gaps" or discontinuous trigrams, and one can expect to locate these in a smaller corpus than full trigrams with the same three elements. We have shown [33] that the whole corpus that would give all 3-slot skip grams is much smaller than that for true trigrams, can probably be computed without loss of generality, and from a corpus not much larger than the web now is, But first, let us ask what would be the point in a fuller associationist model like this, one that covered a language, English, say: how could having that get us closer to rebuilding concepts from all this data, on the assumption that that is what we really want to do, and is the key challenge of machine language understanding?

Let me give two simple examples of this, one from Jelinek's own laboratory (REF) where they showed that simple association criteria could determine semantically coherent classes of objects far more easily than had been thought, provided one had enough data. One can see this most easily now on Google, where what was a research discovery fifteen years ago IBM is now a toy. On labs.google.com/sets one can input any small set of objects one likes and ask Google to find more, in response to this request, from the more than 8 billion English pages it indexes. So, if one types in "Scots, Bavarian, American, German", Google replies with something like "French, Chinese, Japanese, etc". In other words, it has "grasped the concept" of nationality words from context and is, as Wittgenstein would put it, able to go on. This is most certainly a derivation of something clearly semantic from nothing but word data, the problem being that the system does not know what the name of the class is!

A second notion is that of ontologies, forms of knowledge representation that have now become the standard way of looking at formalised knowledge in a wide range of AI, science, medicine and web applications: they contain technical and everyday information structured by set inclusion and membership as well as functional, causal etc. information about sets and objects, and they may or may not have additional strong underlying logical structure. The problem about such structures has always been, as with other forms of knowledge representation discussed here, that they are traditionally written down by human intuition. So what are we to make of the meanings of the terms they contain: are they referential or causal in meaning and can we gather anything from looking at their place in an ordered ontological hierarchy?

This is a straightforwardly Wittgensteinian question and the only proper answer is his own: namely that we cannot tell any term's meaning by looking at it, only by seeing it deployed in use. It is a corollary of that view, assumed in this paper, that all such terms are terms in a language, the language they appear to be in (usually English), and that is so no matter how much their designers protest to the contrary. This is an issue discussed in detail in (Nirenburg and Wilks, [34]).

Ontologies, then, pose something of the problem here that logic traditionally does, or do formal features in linguistics (such as Fodor & Katz' semantic markers [35].): they are claimed to be formal objects, kept apart from language and its vagaries, and with only the meanings assigned to them by scientists. But this isolation cannot in fact be maintained (see Mellor's [36] dispute with Putnam on this issue [37]), and a more reasonable position is that ontologies will have justifiable meanings when they can be linked directly to language corpora, chiefly by being built automatically from them, and subsequently maintained automatically consistently with future corpora. An example of such a current project is ABRAXAS [38], one of a number of projects that claims to do exactly that.

Elements of an ontology can be thought of as triples (e.g. hand – PART OF – body) and our earlier references to skip-grams of trigrams hinted at a large amount of empirical research on using such apparently superficial methods to capture large volumes of such "facts" automatically from large corpora. Such methods go back to the very earliest days of NLP (e.g. [39]. Most recently, a new use for structures of this general type has appeared, namely the subject-relation-object triples (called RDF) that are to carry basic knowledge at the bottom level of the Semantic Web [40], the proposed structure intended to encapsulate human knowledge, based on the world wide web we now have, but annotated in a form to display something of a text's meaning so that computers can use the web themselves. This is too large a vision to discuss here, but one last historical association may be worth making.

Bar Hillel [41] famously attacked the very possibility of machine translation (MT) on the ground that the kinds of

interpretation that translators make require knowledge of vast numbers of facts about the world, and machine translation would therefore need them too. So, you cannot interpret (and so translate) "carbon and sodium chloride" unless you know whether or not there is such a thing as carbon chloride, and so know the inner structure of that phrase (i.e. as carbon+sodium chloride versus carbon chloride + sodium chloride – and it is of course the first of those in this universe). Bar Hillel went on to argue that machines could not have such extensive knowledge of the facts of the world, and so MT was demonstrably impossible.

It was from exactly that point, conceptually if not historically, that AI set out on its long journey to develop mechanisms for representing all the facts in the world (of which the CyC project, [42], is the longest running example.) All this was done in a practical spirit, of course, with no thought or memory of Wittgenstein's declaration that the world was the totality of facts (REF), and what if anything that could possibly mean. It was all practical, energetic computation rather than philosophical thinking, but still, in some sense, fell under Longuet-Higgins' famous adaptation of Clausewitz, that AI was the pursuit of metaphysics by other means. It is interesting that empirically-based NLP has now brought back concepts like the derivation of a totality of facts, not painfully hand-constructed as in CyC, but extracted perhaps by relative simple means from the vast resources of the web's corpora.

4 CONCLUSION

The new vision of the Semantic Web (SW)[40] is in part a revival of the traditional AI project to formalise knowledge, but now also a scientific reality in that so much of science and medicine is already encoded, indispensably so, in structures of this general sort. The process of its construction requires giving meaning progressively to the "upper level" concepts in its ontologies. These upper level concepts are still written down by intuition, which may have validity in scientific area if done by experts – who else can write a map of biology? – But is as much at risk as all the knowledge structures in AI if not grounded in something firmer. I want to argue, in conclusion, that the future SW may offer the best place to see the core of a Wittgensteinian computational linguistics coming into being, as a way of grounding high-level concepts, such as the primitives at the tops of ontologies (e.g. neutrinos, Higgs boson, genes), in real usage of the sort we see in the web-as-corpus.

What I think we are seeing in the SW is a growing together of these upper conceptual levels based on the name spaces and RDF triples derived from texts by skip grams or richer techniques like Information Extraction [43], a successful shallow technology for extracting items and facts that now rests wholly on the success of automated annotation. My belief is that the top and bottom levels will grow together and that interpretation or meaning will "trickle up" from the lower levels to the higher: this is the only way one can imagine the higher conceptual labels being justified on an empirical base. It is a process reminiscent of the concept of "semantic ascent" pioneered by Braithwaite [44] as a description of the way in which interpretation "trickled up" scientific theories from observables like cloud-chamber tracks to unobservables like neutrinos. It is hard to imagine any other route from the distributional analysis on which the revolution in language processing rests up to the interpretation of serious scientific concepts. It is also a process reminiscent of Kant's dictum synthesising Rationalism and Empiricism: "Concepts without percepts are empty; percepts without concepts are blind."

I would argue that the SW is a development of great importance to AI as a whole, even though we still dispute about what it means, and how it can come into being. Many seem to believe that it means Good Old Fashioned AI (GOFAI) is back in a new form, a rebranding of the old tasks of logic, inference, agents and knowledge representation. Core AI tasks have come to something of an impasse: we do not see them marketed much in products after fifty years of research. But a key feature of the SW is that its delivery must be gradual, coming into being at points on the World Wide Web (WWW), probably starting with the modelling of biology and medicine. One cannot easily imagine how it could start somewhere completely new, and without being piggy-backed in on the WWW, yet it will be much more than those same texts "annotated with their meanings", as some would put it.

The key possibility I think the SW offers to traditional AI is to deliver some of its value in a depleted form initially, by trading representational expressiveness for tractability, as some have put it. The model here could be search technology and machine translation on the WWW (or even speech technology): each is available now in forms that are not perfect but we cannot imagine living without them. This may all seem obvious, but machine translation has only recently crossed the border from impossible (or failed) to commonplace. It is far better for a field to be thought useful, if a little dim at times, than impossible or failed. It will be important that web services using the Semantic Web are chosen so as not to be crucial, but merely a nuisance, should they fail. My own current interests are in lifelong personal agents, or Companions, conversationalists as well as agents, where it should not matter if they are sometimes wrong or misleading, any more than it does for people.

This view of the future of the SW is personal and partial; many researchers do not see the need to justify the meanings of logical predicates or ontological terms any more now than they did when they set out in AI and representation in the Sixties. But the history of the CyC project is a good demonstration, if one were needed, of why that cannot be a foundation for AI in the long term: in that project it has not proved possible to keep the interpretation of logical predicates stable over the decades of the system's development; this is a highly significant long-term experimental result for those who believe in the immutability of the meanings of formal items. There is a related view, also current in the SW, that meanings will be saved or preserved by trusted data bases of objects (URIs), referential items in the world, rather in the way digit strings "ground" personal phone numbers in a data base. But this way out will not protect knowledge structures from the changes and vagueness of real words in use by human beings. Putnam considered this problem in the Sixties and declared that scientists should therefore be the ultimate "guardians of meaning". As long as they knew what "heavy water" really meant, it did not matter whether the public knew and perhaps better if they did not. But people call heavy water "water" because it is – because it is indistinguishable from water – otherwise it would have been called "deuterium dioxide". We, the people, are the guardians of meaning and "getting meaning into the machine", probably via the SW, should entail doing it our way, and what could be more in the spirit of Wittgenstein than that?

REFERENCES

[1] Y. Wilks. The History of Natural Language Processing and Machine Translation. In *Encyclopedia of Language and Linguistics,* Kluwer: Amsterdam. (2005).

[2] N. Chomsky. *Syntactic Structures*, Mouton: The Hague. (1957)

[3] R. Carnap, *Logische Syntax der Sprache. English* translation 1937, The Logical Syntax of Language. Kegan Paul, London. (1936)

[4] C. H. Brown, *Wittgensteinian Linguistics.* The Hague: Mouton & Co., (1974)

[5] B. Malinowski, The problem of meaning in primitive languages. In: C.K. Ogden & I.A. Richards (Eds*.), The meaning of meaning*, pp. 296-346. London: Routledge & Kegan Paul. (1923)

[6] W. V. O. Quine, *Word and Object*, Cambridge, Cambridge UP (1960).

[7] L. Wittgenstein, *Tractatus Logico-philosophicus*, Routledge: London. (1961)

[8] Wittgenstein, L. *Philosophische Untersuchungen, Philosophical Investigations*, 2nd ed. Oxford: Basil Blackwell, (1958)

[9] G. Sampson, G*: The 'Language Instinct' Debate*, Continuum, (2004).N.

[10] N. Chomsky,. (1985) Aspects of the Theory of Syntax. Cambridge: The MIT Press.

[11] J. McCarthy and P.J. Hayes. (1969) Some philosophical problems from the point of view of Artificial intelligence. *In Machine Intelligence 4,* (Eds.) Michie and Meltzer, Edinburgh, Edinburgh UP.

[12] J. Veronis, (1993) Sense tagging, does it make sense? http://citeseer.ist.psu.edu/685898.html

[13] R. Moon, (2007) Sinclair, lexicography, and the Cobuild Project: The application of theory. International Journal of Corpus Linguistics, Volume 12, Number 2, 2007.

[14] K. Church , W. Gale , P. Hanks , D. Hindle, (1989) Parsing, word associations and typical predicate-argument relations,
Proceedings of the workshop on Speech and Natural Language*,*
October 15-18, 1989, Cape Cod, Massachusetts.

[15] M. Masterman, (2006) Language, Cohesion and Form: selected papers*,* (Ed.) Y. Wilks, Cambridge UP, Cambridge.

[16] L. Wittgenstein, (1958) The Blue and Brown Books, Oxford: Basil Blackwell.

[17] A. Kilgarriff, G. Grefenstette (2003). Introduction to the Special Issue on Web as Corpus. .*International Journal of Corpus Linguistics* 6 (1).

[18] G. Grefenstette, (2002) Lecture, Sheffield University.

[19] R. K. Moore, (2007) Spoken language processing: Piecing together the puzzle, Speech Communication, 49.

[20] T. Dunning. (1993) Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics.

[21] A. Krotov, R. Gaizauskas, and Y. Wilks. (2001) Acquiring a stochastic context-free grammar from the Penn Treebank. In Proceedings of Third Conference on the Cognitive Science of Natural Language Processing.

[22] K. Goedel. Ueber formal unentscheidbare S Saetze der Principia Mathematica und verwandter Systeme . In Solomon Feferman,(Rd.)*Kurt Goedel: Collected Works, volume 1*, pages 144–195. Oxford University Press, German text, parallel English translation. (1986)

[23] Wittgenstein, L*. Remarks on the Foundations of Mathematics*, rev. edn, ed. G. H. von Wright, R. Rhees, and G.
M. Anscombe, trans. G. E. M. Anscombe, Cambridge, MA: MIT Press.(1978)

[24] Y. Wilks, Y. Decidability and Natural Language, *Mind* LXXX (1971).

[26] Y. Wilks, Y. Preference Semantics, In *The formal semantics of natural language* (Ed.) E. Keenan, Cambridge, Cambridge UP (1975) [27]C. Fillmore, The Case for Case. In Bach and Harms (Ed.): *Universals in Linguistic Theory*. New York: Holt, Rinehart, and Winston, 1-88. (1968)

[28] C. Fillmore, Frame semantics and the nature of language,In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech.* Volume 280: 20-32.(1976).

[29] P.F. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R.L. Mercer, P. Roossin. A Statistical Approach to Machine Translation. *Computational Linguistics*
16:2: 79-85, (1990)

[30] D. L. Waltz, J. B. Pollack: Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation.
Cognitive Science 9(1): 51-74 (1985)

[31] T. Sejnowski and C. Rosenberg. Parallel networks that learn to pronounce English text. *Complex Systems*, 1:145— 168, (1987).

[32] A. Singhal. Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24 (4).p 35-43.(2001)

[33] D. Guthrie, B. Allison, W. Liu, L. Guthrie, Y. Wilks, Y. A Closer Look at Skip-gram Modelling. In *Proc. Fifth International Conference on Language, Resources and Evaluation* (LREC'06), pp. 1222-1225, (2006).

[34] S. Nirenburg and Y. Wilks. What's in symbol. In *Journal of Theoretical and Experimental AI (JETAI)*(2000)

[35] J.J. Katz and J. Fodor. The structure of a semantic theory, *Language* (1963).

[36] D. H. Mellor, Natural Kinds, *British Journal for the Philosophy of Science* 28 (1977).

[37] H. Putnam, Is Semantics Possible*? Metaphilosophy 1 :* p187-201.(1970)

[38] C. Brewster, H. Alani, S. Dasmahapatra, and Y. Wilks. Data-driven Ontology Evaluation. In *Proc. of 4th International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal. (2004)*

[39] Y. Wilks, *Text Searching with Templates*. Cambridge Language Research Unit Memo, ML.156. (1964)

[40] T. Berners-Lee, J. Hendler, and O. Lassila, The Semantic Web, *Scientific American*, May 2001, p. 29-37.(2001)

[41] Y. Bar Hillel, *Language and Information*. Reading, MA: Addison Wesley. (1964)

[42] D. Lenat and R. V. Guha.. *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project.*
Addison-Wesley. (1990)

[43] http://en.wikipedia.org/wiki/Information_extraction

[44] Braithwaite, R. *Scientific Explanation*, Cambridge UP, Cambridge. (1956)

Cognition without content

Paul Schweizer[1]

Abstract. According to the traditional conception of the mind, semantical content is perhaps the most important feature distinguishing mental from non-mental systems. And this traditional conception has been incorporated into the foundations of contemporary scientific approaches to the mind, insofar as the notion of 'mental representation' is adopted as a primary theoretical device. Symbolic representations are posited as the internal structures that carry the information utilized by intelligent systems, and they also comprise the formal elements over which cognitive computations are performed. But a fatal tension is built into the picture - to the extent that symbolic 'representations' are formal elements of computation, their alleged content is completely gratuitous. I argue that the computational paradigm is thematically inconsistent with the search for content or its supposed vehicles. Instead, the concern of computational models of cognition should be with the *processing structures* that yield the right kinds of input/output profiles, and with how these structures can be implemented in the brain.

1 CLASSICISM

According to the traditional conception of the mind, semantical content is perhaps the most important feature distinguishing mental from non-mental systems. For example, in the scholastic tradition revived by Brentano [1], the *essential* feature of mental states is their 'aboutness' or intrinsic representational aspect. And this traditional conception has been incorporated into the foundations of contemporary scientific approaches to the mind, insofar as the notion of 'mental representation' is adopted as a primary theoretical device. For example, in classical (e.g. Fodorian) cognitive science, Brentano's legacy is preserved in the view that the properly cognitive level is distinguished precisely by appeal to representational content. There are many different levels of description and explanation in the natural world, from quarks all the way to quasars, and according to Fodor, it is only when the states of a system are treated as representational that we are dealing with the genuinely cognitive level.

The classical paradigm in cognitive science derives from Turing's basic model of computation as rule governed transformations on a set of syntactical elements, and it has taken perhaps its most literal form of expression in terms of Fodor's Language of Thought hypothesis (henceforward LOT) [2], wherein mental processes are explicitly viewed as formal operations on a linguistically structured system of internal symbols. In particular, propositional attitude states, such as belief and desire, are treated as computational relations to sentences in an internal processing language, and where the LOT sentence serves to represent the propositional content of the intentional state. Symbolic representations are thus posited as the internal structures that carry the information utilized by intelligent systems, and they also comprise the formal elements over which cognitive computations are performed. According to the traditional and widely accepted belief-desire framework of psychological explanation, an agent's actions are both *caused* and explained by intentional states such as belief and desire. And on the LOT model, these states are sustained via sentences in the head that are formally manipulated by the cognitive processes which lead to actions.

Fodor notes that particular tokens of these LOT sentences could well turn out to be specific neuronal processes or brain states. The formal syntax of LOT thus plays a crucial triad of roles: it can represent meaning, it's the medium of cognitive computation, and it can be physically realized. So the syntax of LOT can in principle supply a link between the high level intentional description of a cognitive agent, and the actual neuronal process that enjoy causal power. This triad of roles allows content bearing states, such as propositional attitudes, to explain salient pieces of behavior, such as bodily motions, if the intermediary syntax is seen as realized in neurophysiological configurations of the brain. Because the tokens of LOT are semantically interpretable and physically realizable, they form a key theoretical bridge between content and causation. In this manner, a very elegant (possible) answer is supplied to the longstanding theoretical question of how mental states, such as beliefs and desires, could be viewed as causes of actual behaviour, without violating fundamental conservation laws in physics.

So at first sight, this computational approach to cognition might seem to provide a compelling and harmonious theory of the mind/brain, potentially uniting the traditional notion of mental representation with the causally efficacious level of neural machinery. But alas, a fatal tension is already built into the picture: a central purpose of the symbolic structures is to carry content, and yet, to the extent that they are formal elements of computation, their alleged content is completely gratuitous. Computation is essentially a series of manipulations performed on *uninterpreted* syntax, and formal structure alone is sufficient for all effective procedures. The specification and operation of such procedures makes no reference whatever to the intended meaning of the symbols involved. Indeed, it is precisely this limitation to syntactic *form* that has enabled computation to emerge as a mathematically rigorous discipline. If syntax alone is not sufficient, and additional understanding or interpretation is required, then the procedure in question is, by definition, *not* an effective one. But then the purported content of mental 'representations' is rendered superfluous to the computations that comprise the 'cognitive' processes of cognitive science. The intended interpretation of internal syntax makes absolutely no difference to the formal mechanics of mind.

2 CONNECTIONISM

For a number of years now there has been a high profile struggle between opposing camps within the computational approach to the mind. In contrast to the classical paradigm derived from Turing, connectionist systems are based on networks of large numbers of simple but highly interconnected units that are brain-

[1] School of Informatics, Univ. of Edinburgh, EH8 9LE, UK. Email: paul@inf.ed.ac.uk.

[2]

like in their inspiration. But according to Fodor [3], the brain-like architecture of connectionist networks tells us nothing about their suitability as models of *cognitive* processing, since it still leaves open the question of whether the mind is such a network at the representational level. He concedes that the connectionist approach may be the right type of architecture for the medium of implementation, which would mean that it characterizes a level below that of genuine mental structure. In view of the foregoing tension within the classical paradigm concerning formal syntax and the ineffacacy of content, I would argue that Fodor is on very weak ground when he insists that, within a computational approach, the representational level is fundamental. However, a number of connectionists have taken up the challenge and seek out ways of projecting representational content onto artificial neural networks.

One comparatively recent such attempt (Churchland [4], Laakso, A. and G. Cottrell [5], O'Brien, G. and J. Opie [6]) uses cluster analysis to locate 'vehicles' of representational content within artificial neural networks, where such clusters serve as surrogates for the classical notion of internal syntax. Along with serious difficulties in equating clusters with the syntax of traditional computation, I would contend that such attempts suffer from exactly the same built-in tension that afflicts the LOT model; namely, the purported content for which the clusters serve as vehicles does no work in the processing path leading from inputs to outputs. Just as in the classical case, the postulation of content within the connectionist framework is gratuitous, because it plays no role in the cognitive manipulation of inputs to yield the salient outputs. Indeed, if content weren't gratuitous, then computational versions of cognitive processing would be lamentably deficient in terms of their specification of the inputs. These are characterized solely in formal or syntactical terms, and content is entirely absent from the external stimuli recognized by the operations that can be defined within the model. If representational content were at all relevant, then cognitive systems would have to process content *itself*. But according to computational methods, content is not specified with the input, nor does it play any efficacious role in internal processing. So, from a perspective that takes computation as the theoretical foundation for cognition, it seems quite retrograde to posit content on top of the factors that do the actual work. Surely this is an exemplary occasion for invoking Ockham's razor.

3 THE CHINESE ROOM

Of course, John Searle's celebrated Chinese Room Argument (henceforward CRA) [7] runs the dialectic in exactly the reverse direction: rather than taking the formal, syntactic nature of computation as a reason for eschewing content in a properly naturalistic approach to the mind, Searle instead takes it as a reason for rejecting computation as the appropriate theory of the mental.

So, from the perspective of the present discussion, it is instructive to explicitly cast Searle's argument in terms of the separability of syntactical structure from its intended meaning. In what follows I will abstract away from the somewhat picturesque details of Searle's original version and express the logical core of the CRA via two premises and a conclusion:

> (1) semantical content is an essential feature of the mind,
> (2) syntactical manipulations cannot capture this content, therefore
> (3) the mind cannot be reduced to a system of syntactical manipulations.

Preimse (1) is an expression of the traditional conception of mentality, and is accepted by both Searle and by his opponents in orthodox cognitive science and AI. As stated above, classical cognitive science and AI view the mind according to the model of rule governed symbol manipulation, and premise (1) is embraced insofar as the manipulated symbols are supposed to possess representational content. Searle's dispute with cognitive science and AI centers on his rejection of the idea that internal computation can shed any real light on mental content, which leads to his conclusion (3), and to a concomitant dismissal of the research paradigm central to cognitive science and AI.

In response, a standard line for defenders of the paradigm is to try and defuse the CRA by arguing against premise (2), and claiming that the manipulated symbols really do possess some canonical meaning or privileged interpretation. However, I would urge that this is a strategic error for those who wish to defend the computational approach. As stated above, a distinguishing mathematical virtue of computational systems is precisely the fact that the formal calculus can be executed without any appeal to meaning. Not only is an interpretation intrinsically unnecessary to the operation of computational procedures, but furthermore, there is no unique interpretation determined by the computational syntax, and in general there are arbitrarily many distinct models for any given formal system.

Many classical *negative* results in mathematical logic stem from this separability between formal syntax and meaning. The various upward and downward Löwenheim-Skolem theorems show that formal systems cannot capture intended meaning with respect to infinite cardinalities. As another eminent example, Gödel's incompleteness results involve taking a formal system designed to be 'about' the natural numbers, and systematically reinterpreting it in terms of its own syntax and proof structure. As a consequence of this 'unintended' interpretation, Gödel is able to prove that arithmetical truth, an exemplary *semantical* notion, cannot, in principle, be captured by finitary proof-theoretic means.

Computational formalisms are syntactically closed systems, and in this regard it is fitting to view them in narrow or solipsistic terms. They are, by their very nature, independent of the 'external world' of their intended meaning and, as mentioned above, they are incapable of capturing a unique interpretation, since they cannot distinguish between any number of alternative models. This can be encapsulated in the observation that the relation between syntax and semantics is fundamentally *one-to-many*; any given formal system will have arbitrarily many different interpretations. And this intrinsically one-to-many character obviates the possibility of deriving or even attributing a unique semantical content merely on the basis of computational structure.

These (and a host of other) powerful results on the inherent limitations of syntactical methods would seem to cast a rather deflationary light on the project of explicating *mental content* within a computational framework. Indeed, they would seem to render such goals as providing a computational account of natural language semantics or propositional attitude states profoundly problematic. Non-standard models exist even for such rigorously defined domains as first-order arithmetic and fully axiomatized geometry. And if the precise, artificial system of first-order arithmetic cannot even impose isomorphism on its various models, how then could a *program*, designed to process a specific natural language, say Chinese, supply a basis for the claim that the units of Chinese syntax posses a *unique* meaning?

So I think that the advocates of computationalism make the wrong move by accepting Searle's bait and taking on board the seemingly intransigent 'symbol grounding problem' that results. Instead I would accept Searle's negative premise (2) and agree that computation is too weak to underwrite any interesting version of (1). Hence I would concur with Searle's reasoning to the extent of accepting the salient *conditional* claim that *if* (1) is true *then* (3) is true as well. So the real crux of the issue lies in the truth-value of (1), without which the consequent of the *if-then* statement cannot be detached as a free-standing conclusion. Only by accepting the traditional, *a priori* notion of mentality assumed in premise (1), does (3) follow from the truth of (2). And it's here that I would diverge from the views of both Searle and orthodox cognitive science.

4 CONSCIOUS PRESENTATION

In explicating and defending his pivotal premise (1), Searle [8, 9] again follows Brentano, in claiming that the human mind possesses original intentionality because it can experience conscious presentations of the objects that its representational states are 'about'. Thus it is conscious experience that ultimately underwrites the intrinsic aboutness of genuine intentional states. So Searle holds that consciousness supplies the basis for the truth of premise (1), and he further believes that consciousness arises from the specific causal powers of the brain considered as a physical structure, rather than from multiply realizable symbol manipulation. Hence intentionality is tethered to brain processes via consciousness, and Searle thereby attempts to naturalize the traditional notion of mentality, while at the same time discrediting the computational paradigm, since he argues that computation has nothing to do with consciousness.

And while I would agree with Searle's view that consciousness arises from physical brain activities rather than from multiply realizable computational structure, I would nevertheless argue, contra Searle, that conscious experience, just like symbol manipulation, is too weak to underwrite any interesting version of tenet (1). With respect to the view that conscious experience is the cornerstone of intentionality, the CRA simply begs the question, because it presupposes that the homunculus Searle, replete with conscious presentations, *really does* understand English in some special way. Searle appeals to himself as the locus of genuine intentionality in the Chinese Room, and he would support this by citing the fact that he is consciously aware of the meanings of English expressions. For example, he can entertain a conscious image of the referent of the English string 'h-a-m-b-u-r-g-e-r', while for him the strings of Chinese characters are completely devoid of conscious meanings. Ostensibly, this special understanding of English enables him to follow the program and manipulate the 'meaningless' Chinese symbols. Hence lack of conscious presentation with respect to the semantics of Chinese constitutes the real asymmetry between the two languages, and this underlies Searle's claim that genuine understanding occurs in the case of one language and not the other.

But this line of thought is not particularly compelling, since one can easily concede that Searle has episodes of conscious awareness which attend his processing of English, while at the same time denying that these episodes are sufficient to establish intrinsic content, or to ground the semantics of natural language expressions. Indeed, the mere occurrence of conscious presentations is too weak to even establish that they themselves play a role in Searle's ability to follow the English instruction manual. Instead, I would argue that what consciousness actually provides is the foundation for the subjective *impression*, had by Searle and others, that the human mind enjoys some mysterious and seemingly magical form of intentionality with the power to uniquely determine representational content.

Thus when Searle contends that our mental states are 'really about' various external objects and states of affairs, this is merely an expression of the fact that, introspectively, it *seems to us* as if our mental states had some such special property. Conscious experience is clearly sufficient to provide the source for this belief, since conscious experience determines how (some of) our mental states appear to us. But it cannot provide a basis for concluding that the belief is *true*, unless consciousness is something much more mysterious and powerful than the resources of natural science can allow. Brentano famously dismissed naturalism, and he thereby gave himself some room for the claim that consciousness underwrites the mind's essential intentionality. However, if one accepts naturalism and views consciousness as a phenomenon supported by, say, the causal properties of electrochemical reactions taking place inside the skull, then one should just bite the bullet and accept that it is too weak to support Brentano's thesis that intentionality is an essential feature of the mind.

It would be straying too far from the main goal of the article to expand on this latter claim at any great length, but considerations based on the 'narrow' status of consciousness should suffice to illustrate the central point. It is widely held by naturalists that occurrent conscious states must supervene upon occurrent, *internal*, physical states and processes of organisms. As a consequence, something outside the boundaries of an organism cannot affect consciousness, unless it makes some relevant impact on the occurrent, internal physical states and processes, most typically through inputs to the sensory mechanisms. But then the objection raised by Searle in the CRA against the computational paradigm comes back to undermine his own position: the relation between consciousness and its object becomes one-to-many, just as the relation between computational syntax and its interpretation is one-to-many. Any number of different external causes could yield exactly the same conscious experience (by inducing exactly the same internal physical states and processes), just as a given formal system can have arbitrarily many distinct interpretations. Therefore conscious experience is, by its very nature, too weak to determine a unique object that one is conscious of. This problem is at the heart of Cartesian scepticism, and it only gets worse within the narrow confines of naturalism. In a more contemporary guise, Putnam's celebrated brains-in-a-vat argument [10] exploits this solipsistic feature to show that conscious psychological states are too weak to capture the semantics of natural language.

5 ANTI-REPRESENTATIONALISM

There have been a number of high profile positions advanced in negative reaction to 'classical' cognitive science that take anti-representationalism as one their hallmarks, including dynamical systems theory (e.g Van Gelder [11]), behaviour based robotics (e.g. Brooks [12]), approaches utilizing sensory-motor affordances (e.g. Noë [13]), and some varieties of connectionism. A common factor is that these views all advance some version of the slogan 'intelligence without representation'. In order to locate my position on the salient philosophical landscape, it is worth noting that it is *not* anti-representational in this sense. On my view, there could well be internal structures that play many of the roles that people would ordinarily expect of representations, and

this is especially true at the level of perception, sensory-motor control and navigation. So I would be quite happy to accept things like spatial encodings, somatic emulators, internal mirrorings of relevant aspects of the external environment. Ultimately this boils down to questions that have to be settled empirically in the case of biologically induced agents, but unlike the anti-representationalists, I do not deny that the most plausible form of cognitive architecture may well incorporate internal structures and stand-ins that many people would be tempted to *call* 'representations'.

But I would argue that this label should be construed purely in a weak, operational sense, and should not be conflated with the more robust traditional conception. To the extent that internal structures can encode, mirror or model external objects and states of affairs, they do so via their own causal and/or syntactic properties. And again, to the extent that they influence behaviour or the internal processing of inputs to yield outputs, they do this solely in virtue of their causal and/or syntactic properties. There is nothing about these internal structures that could support Searle's or Brentano's notion of original intentionality, and there is no independent or objective fact of the matter regarding their 'real' content or meaning.

So what I deny is not that there may be internal mechanisms that reflect external properties in various systematic and biologically useful ways. Instead I would deny that there is anything *more* to this phenomenon than highly sensitive and evolved relations of calibration between the internal workings of an organism and its specialized environmental context. Evolutionary history can be invoked to yield interesting heuristics with respect to these physical relations of calibration, and perhaps support counterfactuals regarding their role in the organism's adaptive success. But evolution is based on random mutation, and natural 'selection' is an equally purposeless mechanism. Neither can provide the theoretical resources sufficient to ground the strong traditional notion of 'genuine aboutness'.

Thus if I had to coin a competing slogan to encapsulate my own position, it would be something like 'representation without intentionality'. If one is truly committed to naturalism, then there is only a difference of degree and complexity, but not in kind between, say, the reflection of moonlight in a pond and the retinal image of the moon in some organism's visual system. Proponents of the orthodox view are inclined to think that a sufficient difference in degree and complexity somehow yields an esoteric difference in *kind*, a difference that allows us to cross the conceptual boundary from mere causal correlations to 'genuine aboutness'. But I would contend that naturalism itself supplies an asymptotic limit for this curve, and that the boundary can be crossed only by invoking non-natural factors.

6 CONCLUSION

According to the position advocated herein, Fodor's characteristic insistence on representational *content* embodies an unfortunate commitment to an a priori view of the mind that does not fit within the context of naturalistic explanation. The crucial point to notice is that internal 'representations' do all their scientifically tangible *cognitive* work solely in virtue of their physical/formal/mathematical structure. There is nothing about them, qua efficacious elements of internal processing, that is 'about' anything else. Content is not an explicit component of the input, nor is it acted upon or transformed via cognitive computations. All that is explicitly present and causally relevant are computational structure plus supporting physical mechanisms, which is exactly what one would expect from a naturalistic account.

In order for cognitive structures to do their job, there is no need to posit some additional 'content', 'semantical value', or 'external referent'. Such representation talk may serve a useful heuristic role, but it remains a conventional, observer-relative ascription, and accordingly there's no independent fact of the matter, and so there isn't a sense in which it's possible to go wrong or be mistaken about what an internal configuration is 'really' about. Instead, representational content can projected onto an internal structure when this type of gloss plays an opportune role in characterizing the overall processing activities which govern the system's interactions with its environment, and hence in predicting its salient input/output patterns. But it is simply a matter of convenience, convention and choice, and does not reveal an underlying fact of the matter nor any essential characteristics of the system.

From the point of view of the system, these internal structures are manipulated *directly*, and the notion that they are 'directed towards' something else plays no role in the pathways leading from cognitive inputs to intelligent outputs. Hence the symbol grounding problem is a red herring – it isn't necessary to quest after some elusive and mysterious layer of 'real' content, for which these internal structures serve as the mere syntactic vehicle. Syntactical and physical processes are all we have, and their efficacy is not affected by the purported presence or absence of meaning. I would argue that the computational paradigm is thematically inconsistent with the search for content or its supposed vehicles. Instead, the concern of computational models of cognition should be with the internal *processing structures* that yield the right kinds of input/output profiles of a system embedded in a particular environmental context, and with how such processing structures are implemented in the system's physical machinery. These are the factors that do the work and are sufficient to explain all of the empirical data, and they do this using the normal theoretical resources of natural science. Indeed, the postulation of content as the essential feature distinguishing mental from non-mental systems should be seen as the last remaining vestige of Cartesian dualism, and, contra Fodor, naturalized cognition has no place for a semantical 'ghost in the machine'. When it comes to computation and content, only the vehicle is required, not the excess baggage.

REFERENCES

[1] F. Brentano, *Psychology from an Empirical Standpoint*. (1874).

[2] J. Fodor, *The Language of Thought*, Harvester Press, (1975).

[3] J. Fodor and Z. Pylyshyn, Connectionism and Cognitive Architecture: A Critical Analysis, *Cognition,* 28: 3-71, (1988).

[4] P. M. Churchland, Conceptual Similarity Across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered, *Journal of Philosophy*, 96(1): 5-32, (1998).

[5] A. Laakso and G. Cottrell, Content and Cluster Analysis: Assessing Representational Similarity in Neural Systems, *Philosophical Psychology*, 13(1): 47-76, (2000).

[6] G. O'Brien and J. Opie, Connectionist Vehicles, Structural Resemblance, and the Phenomenal Mind, *Communication and Cognition*, 34: 13-38, (2001).

[7] J. Searle, Minds, Brains and Programs, *Behavioral and Brain Sciences*, 3: 417-424, (1980).

[8] J. Searle, Consciousness, Explanatory Inversioin and Cognitive Science, *Behavioral and Brain Sciences*, 13: 585-596, (1990).

[9] J. Searle, *The Rediscovery of the Mind*, MIT Press, (1992).

[10] H. Putnam, Brains in a Vat. In: *Reason, Truth and History*, H. Putnam, Cambridge University Press, (1981).
[11] T. Van Gelder, Dynamics and Cognition. In: *Mind Design II*, J. Haugeland (Ed.), MIT Press, (1996).
[12] R. Brooks, Intelligence without Representation. In: *Mind Design II*, J. Haugeland (Ed.), MIT Press, (1996).
[13] A. Noë, *Action in Perception*, MIT Press, (2004).

Foundations of a Philosophy of Collective Intelligence

Harry Halpin [1]

Abstract. Philosophy, artificial intelligence and cognitive science have long been dominated by the presupposition that intelligence is fundamentally individual. Recent work in cognitive science clearly undermines that notion. Increasingly, intelligence is seen not as having its locus in the individual, but in the network of relationships that the individual has with the external world and other individuals. At the same time, there has been an increasing neo-Heideggerian focus on the role of embodiment and anti-representationalism, as shown by work ranging from robotics to dynamical systems. While philosophers are carefully trying to justify this development, the most significant computational phenomenon by far - the World Wide Web - is a veritable explosion of representations. In its latest stage, the Web has become increasingly more the realm of representations used for social real-time co-ordination, as a tool for "collective intelligence." In order to make sense of these developments, we first summarize the differences between the Cartesian assumptions of classical artificial intelligence and the neo-Heideggerian embodied cognitive science. Then we show both how Brian Cantwell Smith's story of representations can be built on top of a neo-Heideggerian story. A combination of a refined version of Smith's rehabilitation of representationalism with the Extended Mind Hypothesis can explain the emergence of collective intelligence and its mediation through representations, and so the wide-scale success of the Web. Finally, we reconsider the notions of autopoiesis, the individual body and embodiment itself in light of collective intelligence.

1 The Individual Challenged

The paradigmatic problem of both analytic philosophy and cognitive science is to explain the intelligence of the human individual: What properties of the individual human deserve credit for intelligence, and why? The answers seem to be self-evident; the unique combination of language and consciousness of the individual is the foundation of intelligence, both of which are not obviously found in ants, trees, or computers. Language and consciousness both seem to be incarnations of a reasoning process that leads to flexible, adaptive behavior, the general purpose reasoning mechanism of Descartes. Ranging from Frege and Russell onwards, philosophy of language sought to explain the relationship of the logical grammar and the world in order to explain why language is so effective, while more recently philosophers have been flocking to the rather mysterious "hard" problem of consciousness. On a more empirical vein, artificial intelligence attempts to understand intelligence through building mechanisms that display intelligence. Yet after the failure of classical artificial intelligence[2] to produce intelligence in computers that could scale out of very small domains, a strain of research based primarily in robotics have shown that the very details of the implementation can produce intelligent behavior without representations, much less consciousness or reasoning [3]. This empirically-driven focus on embodiment has signalled the greatest change in artificial intelligence since its inception, and is explained by Wheeler as a shift from a classical Cartesian paradigm to a neo-Heideggerian programme [32]. Despite this revolution, one assumption that analytic philosophy, classical AI, and the new embodied AI all share is that the fundamental unit of analysis should be the individual.

Recent empirical work in psychology and cognitive science has increasingly challenged the assumption that intelligence is irreducibly individual. It has shown that for complex tasks such as ship navigation that the success of the action relies on the co-ordination of multiple individuals [16]. Evidence from decision-making shows that the "wisdom of crowds" - in other words, decision-making guided by the aggregate of information in a social network - reliably makes better decisions than any individual [4]. Furthermore, work in developmental psychology has shown that the ability to point in children is more than an expression of a linguistic demonstrative, but rather an effort to produce a shared intentionality by directing the attention of others to the same object [31]. Some evidence from neuroscience the explosion of frontal cortex, long thought to be the seat of reasoning, evolved to keep track of interactions within a social network [8], and that the presence of mirror neurons provides a set of neurological mechanisms that allow individuals to share the same neurological state [9]). More recent work in tracking the behavior of individuals finds that their behavior - ranging from movement to turn-taking in conversation - can be reliably tracked by appealing to the behavior of others in their social network with a high degree of accuracy (over 40 to 80% of variation over a wide variety of tasks) without any appeal to planning, reasoning, or verbal language [24]. Pentland claims "that important parts of our personal cognitive processes are caused by the network via unconscious and automatic processes such as signaling and imitation, and that consequently, important parts of our intelligence depend upon network properties." Instead of locating the intelligence in the individual, intelligence can be located in the collective aggregate of individuals.

Collective intelligence does not necessarily mean the sharing of a cognitive state by, for example, mirror neurons. Intelligence can be exhibited by a network of individuals where each individual is specialized in a particular task so that no two individuals share the same cognitive state (skills, activity, and so on) per se, but that the successful action depends on the activities of the entire network. The classic cognitive ethnographic example by Hutchins is the piloting of a ship, where correct piloting of the ship depends on each individual, ranging from the navigator to the steersmen, completing their task [16]. Furthermore, it is not the simple aggregate or organization of individuals in a network that deserves credit for intelligence, but the

[1] University of Edinburgh email: H.Halpin@ed.ac.uk

[2] It should be noted that artificial intelligence is usually the application of reigning theories in philosophy, and classical artificial intelligence was based on the "Language of Thought" representationalism in philosophy of language [5].

conjunction of this social network with their environment. The environment should not be considered static, but dynamically shaped by the actions of intelligent behavior. However, some of the knowledge needed for success is not just embodied in individuals, but embodied in the environment, in their artifacts such as compasses and maps, and the very shape of the boat itself. This leads us to consider the example put forward by Herbert Simon of the apparent complexity of an ant's path as it steadily marches towards food on the beach: "Viewed as a geometric figure, the ant's path is irregular, complex, and hard to describe. But its complexity is really a complexity in the surface of the beach, not the complexity in the ant" [27]. Although this may be true in some cases, it would be to primitive to describe the ant totally to be at the mercy of its environment. Intelligence in general - collective or not - leave traces behind in the environment. The classic example is the pheromone trace of the ant, in which a traces get reinforced as more ants use a particular trail, has been shown to be an efficient way of navigating the environment. This shows how individuals with limited memory can use the shaping of the environment as an external memory. Culture, ranging from design of cities to Wikipedia, can be considered collective cognition extended into the environment. This usage of the environment has a number of advantages over direct individual-to-individual communication. As noted by Heylighen, there is no need for simultaneous presence, so interaction can be asynchronous, and individuals can even be anonymous and unaware of each other. This allows highly organized successful actions to be performed by individual that, due to limited memory and knowledge, would be unable to achieve success otherwise [14].

To modify Pentland's thesis: The collective activity of individuals and their modifications to the environment are responsible for intelligence. While at first this thesis seems intuitive, it goes against much of the practice of both classical cognitive science and philosophy that have a tendency towards individualist reductionism. While the question of whether or not this thesis is actually true is a distinctly empirical question, the philosophical ramifications of this thesis should be developed to see if they are in conflict or continuous with the neo-Heideggerian framework currently being championed in philosophy and AI. Two points of conflict immediately become apparent. Although Heidegger himself is unclear, the neo-Heideggerian framework as articulated by Wheeler understands intelligence as a function of the situated being in the world, not a collective of beings in a shared world [32]. Furthermore, the neo-Heideggerian framework does not explain the reshaping of the environment by intelligence, in particular the creation of representations, not just representational explanation. Representations are seen as crucial by many for the emergence of collective intelligence, which Hutchins traces his "distributed cognition" to "the propagation of representational states across representational media" [16]. The neo-Heideggerian framework is most associated with robotics that exhibits "intelligence without representation," and in contrast collective intelligence is most associated with the advent of the Web, a veritable explosion of representations if ever there was one.

To tackle these problems, we will focus on them in reverse order. First, after explaining the rising neo-Heideggerian framework in cognitive science by contrasting it with the classical Cartesian framework, we will show how representations can be built into the framework. Then, by pushing on the Extended Mind thesis, we will show how the neo-Heideggerian framework allows collective intelligence, including those that use representations. We can then use this framework to understand the explosion of collective intelligence on the representation-heavy Web, and finally try to reconstruct a notion of what should replace the individual in philosophy.

2 Neo-Heideggerian Embodiment

The philosophical assertions made by proponents of neo-Heideggerian programme must be summarized in order to see if they are continuous, or in contradiction with, a theory of representation-based collective intelligence. This is difficult, as like classical artificial intelligence, the move towards embodiment in AI has mainly been one of empirical work where the philosophical assumptions have for the most part been implicit in the work itself. Just as Dreyfus unearthed the philosophical presuppositions of Cartesian classical artificial intelligence, Wheeler has effectively summarized the assertions of embodied AI and based them firmly on a reading of Heidegger, which we call the *neo-Heideggerian programme* [32]. The neo-Heideggerian programme is best understood in contrast with the neo-Cartesian programme of classical AI. Wheeler digests this programme into three main assumptions:

- The subject-object dichotomy is a primary characteristic of the cognizers ordinary epistemic situation
- Mind, cognition, and intelligence are to be explained in terms of representational states and the ways in which such states are manipulated and transformed.
- The bulk of intelligent human action is the outcome of general purpose reasoning processes that work by retrieving just those mental representations that are relevant to the present behavioral context and manipulating and transforming those representations in appropriate ways as to determine what to do

It should be noted that at first glance these neo-Cartesian assumptions are based on the individual being the locus of intelligence. That is surely how at least Descartes thought of it: The singular subject is operative in "cogito ergo sum." The first of the Cartesian points seems to have an implicit individual subject, while the second remains neutral, and the third also seems to have an implicit human individual as the subject. Wheeler then makes the fairly accurate assessment that "word on the cognitive-scientific street is that classical systems have, by and large, failed to capture in anything like a compelling way, specific styles of thinking at which most humans naturally excel" [32]. However, all hope is not lost for AI if it can only lose its neo-Cartesian assumptions. Based on a survey of current work in AI, ranging across robotics, artificial life, and dynamical systems, Wheeler unifies these diverse works on four new assertions, which he states as follows [32]:

- **The primacy of online intelligence**: The primary expression of biological intelligence, even in humans, consists not in doing math or logic, but in the capacity to exhibit...online intelligence...a suite of fluid and flexible real-time adaptive responses to incoming sensory stimuli.
- **Online intelligence is generated through complex causal interactions in an extended brain-body-environment system**: Online intelligent action is grounded not in the activity of neural states and processes alone, but rather in the complex causal interactions involving not only neural factors, but also additional factors located in the non-neural body and the environment.
- **An increased level of biological sensitivity**: Humans and animals are biological systems - and that matters for cognitive science.
- **A dynamical systems perspective**: Cognitive processing is fundamentally a matter of state space evolution in certain kinds of dynamical systems.

Is there any bias towards an individual subject in these assertions? It seems present in a subtle manner in the first assertion since the

very idea of "incoming sensory stimuli" presumes an individual that is processing these stimuli. The second and third assertion also seem to take for granted that our primary subject is not just an individual, but a biological individual. This is put into perspective by the second assertion that "not only neural factors, but also additional factors located in the non-neural body and the environment" play a critical role, a point we will return to with a vengeance.

Wheeler and his philosophical fellow-travellers such as Clark [5] spend much of their time on the question of whether or not there is any room whatsoever for internal representations inside these individuals. Rejecting Clark's notion of "decoupling" as sufficient but not necessary for cases he believes demands a representational explanation, Wheeler argues for some, albeit limited role for representations that pins representations on the two notions of homuncularity and arbitrariness. Since it is too involved to argue over homuncularity and arbitrariness here, we shall instead focus on how Brian Cantwell Smith's revival of decoupability can be built on a neo-Heideggerian framework. We shall just comment that Wheeler's general framework is not incompatible with our notion of collective intelligence and his account of representations is not too far from our account.

3 Representations Revisited

The very idea of representation is often left under-defined and is as a consequence given near-magical powers by certain theories of language and classical AI. While it is hard to pin down a reigning definition, the classic definition stems from the notion of a "symbol" given by Simon and Newell's *Physical Symbol Systems Hypothesis* [22]:

"An entity X designates an entity Y relative to a process P, if, when P takes X as input, its behavior depends on Y."

First, the very idea of "being a representation" is grounded in the behavior of a process, and behavior depends on having access to the representation. Thus, the target of representation (i.e. what is represented, the "thing designated") will depend on the process the representation is used in, i.e. a representation is never context-free. Second, there is clearly decoupling "for this is the symbolic aspect, that having X (the symbol) is tantamount to having Y (the thing designated) for the purposes of process P" [22]. This definition seems to have an obvious point of conflict with the neo-Heideggerian agenda, for it reflects the infamous "subject-object dichotomy" due to its presupposition of at least three distinct a priori entities, the subject (P), the representation (X), and the object (the "target" of the representation, Y). To the extent that these distinctions are held a priori, then the definition is the very exemplar of the neo-Cartesian programme of classical AI.

An escape-hatch from this Cartesian dead-end would exist if there was a way within the neo-Heideggerian program to tell the story of how representations come to be without an a priori subject-object dichotomy. Brian Cantwell Smith tackles this by developing a theory of representations that does not presume an individual [28]. Smith starts with the example from Lettvin and Maturana, a frog tracking a gadfly across the sky [17]. The frog sees the fly, and begins tracking it with its eyes as it flies. The frog and the gadfly are both physically connected via light-rays. Borrowing an analogy from physics, everything is composed of non-distinct fields of energy, so it would be a presupposition to talk about a frog, a fly and light as individual objects. All that exists is some sort of pre-individual flow from which individual objects may emerge. At the moment of tracking, connected as they are by light, the frog, its light cone, and the fly are a system, not distinct individuals. An alien visitor might even think they were a single individual. When the fly goes behind a tree, and the fly emerges from the other side of the tree, the frog's eyes are not focused on the point the fly was at before it went behind the tree, but the point the fly would be at if it continued on the same path.[3] Components of the flux are now physically separated, with a mutually distinct o-region and s-region. The s-region is distinguished from the o-region by virtue of not only its physical disconnection but by the s-region's attempt to "track" the o-region, "a long-distance coupling against all the laws of physics" [28]. After disconnection (and possibly more cycles of disconnection and re-connection) the s-region can stabilize as an individual subject and the o-region as an individual object, and with considerable work on the subject's side to "track" its object a representation is created by the subject using some form of dynamically incoherent memory. Both subject and object are then full-blown individuals, with the subject possessing a representation of the object[28]. The individuals are not a-priori distinct, but co-constitute each other. According to this explanation subject and objects co-evolve, with the physical processes used to track the object being the representation.

In order to clarify and make abstract Smith's analogy and explicitly connect it to Simon and Newell's definition, we can divide Smith's process into what I have called the *representational cycle* [10]. In order to explicate why precisely the s-region differs from the o-region, we rely on Rocha and Hordijk's work on evolving representations, in particular their idea of dynamically incoherent memory [25]. Dynamically incoherent memory is defined as a type of memory not changed by any dynamic process it initiates or encounters. In this manner, it serves as memory that does not degrade or radically alter, but can maintain itself over time. To phrase this outside of the language of dynamical systems, we would say that "dynamically incoherent" might be a misleading word. Instead, what Rocha means is that the subject must have a some sort of memory that is capable of maintaining coherence in terms of its physical structure against, "the vagaries and vicissitudes, the noise and drift, of earthy existence" as Haugeland would say [11]. The cycle can then be put into four stages [10]:

- **Presentation**: Process S is in effective local contact (i.e. physically in contact in space-time) with process O. S is the s-region that evolves into the subject that has the representation and O is the o-region that evolves into the object.
- **Input**: The process S is in local effective contact with coherent memory R. An input procedure of S puts R in correspondence with some portion of process O. This is entirely non-spooky since S and O are in effective local contact. R evolves into the representation.
- **Separation**: Processes O and S change in such a way that the processes are non-local.
- **Output**: Due to some local effect in process S, S uses its local effective contact with R to initiate the local dynamic behavior that depends on R for success.

Smith, and our exegesis of him, has shown it is possible to build a theory of representations based on decoupability and correspondence

[3] While simple physics can do this without any intentionality by making the frog's eyes continue along at the same trajectory, for more complex behavior, such as when the fly is not moving at a constant rate but zig-zagging about, more complex tracking is required. Regardless, the point of Smith's example is that disconnection is required for decoupability and so representation

while not presupposing that intelligent behavior of an individual cognizer depends on internal representations - or that representations - or even an individual - exist a priori at all. Representations are also not everywhere as in traditional representationalism, but instead they are deployed as needed when the relevant behavior requires distal co-ordination. Representations - if not representationalism - is continuous with the neo-Heideggerian agenda. In fact, the very story of representations gives us a way to show how the notion of an individual can emerge from some primordial and undefined Heraclitan flux. Representations are not a Cartesian metaphysical assumption, but arise over time in even a neo-Heideggerian world.

4 From the Extended Mind to the Web

Now that we have shown a plausible story about how representations can be built on a neo-Heideggerian framework, we have to explain how these representations can be used to explain the rise of a robustly representational system like the Web without contradicting the neo-Heideggerian framework. Once this has been done, we can use the current activity on the Web as pointing the way for questioning our conception of the individual, and thereby questioning the bias towards the individual as the fundamental unit of analysis of even the neo-Heideggerian framework. To return to the task at hand, one principle of the neo-Heideggerian agenda put forward by Wheeler is that "online intelligence is generated through complex causal interaction in an extended brain-body-environment system" [32]. We can press on this assertion to make room for the active role for representations in general, and for the Web in particular, "an active externalism, based on the active role of the environment in driving cognitive processes" [6]. Since Smith's representations are not necessarily internal or external to a process, we can remain agnostic as regards whether or not internal representations are necessary or even used by an individual. For example, a representation can be stored in the memory "inside" the head of an agent in some neural state, but it can just as easily be stored outside in a map. The debate over the existence of internal representations is an empirical debate best left to empirical work. However, what is less debatable seems to be the fact that representations at least exist externally from particular agents. After all, finding those representational neural states are difficult, but let us not deny the existence of maps!

In their Extended Mind Hypothesis, Clark and Chalmers introduce us to Otto, a man with an impaired memory who navigates about his life via the use of notes in his notebook [6]. Otto wants to navigate to the Metropolitan Museum of Modern Art in New York City from his house in Brooklyn, but to do so with his impaired memory he needs at least the address. To specify more than Clark and Chalmers, let us say that he needs a map.[4] In order to arrive at the museum, Otto needs a map whose components are in some correspondence with the world he must navigate in order to get to the museum, in other words a representation. Let us say that Otto has in his notebook a map to the Museum of Modern Art that exists for the precise purpose of navigating individuals to the museum. It is hard to deny that a map is representational in the sense we have presented above, as it is a representation whose target is the various streets on the way to the Museum. The map is just an external representation in the environment of Otto, and can drive the cognitive processes of Otto in a similar fashion to the way that classical AI assumed internal representations in Otto's head did. Clark and Chalmers point out that if external factors are driving the process, then they deserve some of the credit: "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" [6]. In this regard, the Extended Mind thesis undermines the strict division between internal and external of the agent itself, but again, in a way that is compatible with the neo-Heideggerian framework.

Imagine the world to be inhabited by multiple individuals that can access the same representation. In almost all the original examples that Clark and Chalmers use in the Extended Mind argument, they deploy a single person sitting in front of a computer screen [6]. A more intuitive example would be two people using the Internet to both share a single representation. One could imagine Otto trying to find his way to the Museum of Modern Art, and instead of a notebook having a personal digital assistant with access to a map on the Web. Likewise Inga can have access to the exact same map via her personal digital assistant. Since both Otto and Inga are sharing the exact same representation and because they are both using it in the same manner, Inga and Otto can be said to share at least some of the same cognitive state, due to the fact that their individual cognitive states are causally dependent on accessing the same representation. This representation is the "same" precisely because the digital memory of the computer allows "perfect" copies to an extent as Haugeland explains [11]. However, unlike the lone digital computer, what the Web specializes in is allowing *everybody* to access the same set of representations.

The value of external representations comes with their accessibility, for an external representation that is not accessible when its needed cannot be used to enable online intelligence. It is precisely in order to solve this problem that Tim Berners-Lee proposed a World Wide Web as a universal information space [1]. The primary advantage of the Web is that every representation has a unique name, a URI.[5]. The Web allows each representation to be accessed when needed by using its unique name. Combined with the fact that since the representations are digital and can be communicated in a lossless fashion, the Web allows multiple simultaneous accessing of the exact same representation. Since the Web is a universal space of digital representations, two or more individuals can share the same representation simultaneously. Due to the Extended Mind hypothesis, two or more individuals can then, because of simultaneous access, share some of the same cognitive state.

5 The Web as Collective Intelligence

Much as computation has not remained static, neither has the Web. The Web, as originally conceived by its users, was just a collection of documents connected by hyperlinks, albeit one in a universal information space. These documents were mostly static, being authored and maintained by individuals. Although new pages and links could be added without resort to a centralized registry, the content of the Web was for the vast majority of users was not content that they actually created and added to in any meaningful manner. Within the last few years, a combination of easy-to-use interfaces for creating content and a large number of web-sites that prioritize the social and collaborative creation of content by ordinary users have taken off, leading to the phenomenon known as "Web 2.0," literally the next genera-

[4] In fact, many of us would need a map even without an impaired memory, which points to how widespread this phenomenon is.

[5] Originally the "Universal Resource Identifier," now a Uniform Resource Identifier as given in an updated specification [2] These are exemplified by the familiar format of *http://www.example.org*.

tion of the Web.[6]. This transition from the Web of static hyperlinked web-pages to a more interactive and collaborative medium is more accurately described as a transition from a "Web of Documents" to a "Social Web" [15]. Paradigmatic examples of easy-to-use interfaces would be Google Maps (or even Google Earth),[7] while a paradigmatic example of socially-generated content would be Wikipedia[8]. Furthermore, increasingly these web sites are now being woven into the fabric of the everyday life of more and more people. How many people feel that their intelligence in increased when they have immediate access to a search engine to the Web, a massive encyclopedia available in a few seconds notice?

The Social Web then presents an interesting twist on the Extended Mind Hypothesis extension that we presented earlier. Again, Otto is using a web-page in his mobile phone to find his way to the Museum of Modern Art. While our previous example had Otto using the Web as ordinary Web users did years ago, simply downloading some directions and following them, we now add a twist. Imagine not only that Inga and Otto are using a map-producing Web site that allows users to add annotations and corrections, a sort of wiki of maps. Inga, noticing that the main entrance to the Museum of Modern Art is closed temporarily due to construction and so the entrance has moved over a block, adds this annotation to the map, correcting an error as regards where entrance of the Museum of Modern Art should be. This correction is propagated at speeds very close to real-time back to the central database behind the Web site. Otto is running a few minutes behind Inga, and because this correction to the map is being propagated to his map on his personal digital assistant, Otto can successfully navigate to the new entrance a block away. This (near) real-time updating of the representation was crucial for Otto's success. Given his memory issues, Otto would have otherwise walked right into the closed construction area around the old entrance to the Museum and been rather confused. This active manipulation with updating of an external representation lets Inga and Otto possess some form of dynamically-changing collective cognitive state. Furthermore, they can use their ability to update this shared external representation to influence each other for their greater collective success. In this manner, the external representation is clearly social, and the cognitive credit must be spread across not only multiple people, but the representation they use in common to successfully accomplish their behavior. Clark and Chalmers agree, "What about socially extended cognition? Could my mental states be partly constituted by the states of other thinkers? We see no reason why not, in principle" [6]. How we have extended their story is that socially extended cognition is now mediated by external representations, in particular interactive representations on the Web.

Even this example of brings up points for further consideration. Ordinarily as considered in representationalism as a theory of mind, representations are considered notoriously disconnected from their target, and so while this leaves plenty of room to develop a theory of misrepresentation, it leaves quite a lot of work for philosophers to develop how something like an "internal representation" might have a correspondence with a "target" in the external world. Indeed, this understanding of representationalism as some internal language of thought is precisely what we are *not advocating*, for those philosophical problems among others. What our previous example shows is not that representations are some mysterious language of thought, but as Andy Clark put it, "material symbols" capable of being brought into contact with their equally material targets. While the map may not be the territory, it brings Inga and Otto into contact with the territory.

This leads us back full circle to the Web. For example, the collective editing and use of Wikipedia allows its representations to be increasingly part of of the cognitive system of many people. As representations on the Social Web are updated by increasing numbers of people, each representation is increasingly brought into tighter coupling with both its target and the agent using the representation. As each representation is involved in this process of use and updating is brought into closer and closer cognitive updating with more and more individuals, the representations on the Web are brought into tighter and tighter coupling with what its users formerly considered their individual intelligence, and so leading to the phenomenon widely known as collective intelligence. Indeed, there are now problems as simple as navigating down the street or organizing a social event that many today would have difficulty organizing without access to an interactive mapping Web service or a social networking web site. As users contribute more and more content, the collective content of these web-pages becomes increasingly difficult to track down to individuals. Some of these Web-based tools for collective intelligence have no way to track down the original individual author, others like Wikipedia have sophisticated mechanisms in place to track individual contributions. However, as long as the contribution that the collectively-built web page makes is the sum of more than an individual effort, then the credit must be placed upon the collective content, not the individual author. From the standpoint of the user of the representation, the credit must also not just be placed on the creator of the content, but the very technological infrastructure - ranging from the hardware of high-speed fibre optics and wireless routers to the software of protocol design and web server code - that enables the content of the collectively created web site to be delivered when it is needed. The credit for successfully creating and deploying the cognitive scaffolding is more collective than originally thought! It is also this cognitive scaffolding that provides the ability for distributed individuals to rapidly co-ordinate in near real-time through the modifications of representation, so realizing the definition of collective intelligence given by Levy as "A form of universally distributed intelligence, constantly enhanced, coordinated in real time, and resulting in the effective mobilization of skills"[18].

6 Conditions of Collectivity

When one throws even the concept of the a priori individual away, one should seriously reconsider what one is left with. Can we throw away the notion of the individual that just happens to co-incide with what is considered a biological body, the 'common sense' body whose ends happen to coincide with the skin? Obviously upon closer inspection, even the individual biological body is a collectivity, for it is obviously composed of a collective of organs, which are in turn a collective of cells, and so on. If so, should one privilege the biological makeup of certain organs? Evidence from neuroscience in the famous "phantom limb" experiments points to the fact that what our consciousness considers the boundaries of our body does not coincide with our actual biological skin, and that experiments ranging from prosthetic limbs to cochlear implants shows that functionally, non-biological components can be easily considered very much part of the body by the consciousness itself. What we are searching for is then a notion that can define an individual body without resort to making biological tissue some sort of "wonder tissue" as Dennett would put it. One candidate is Maturana's notion of autopoiesis, a more refined notion of the homeostasis that defined earlier cybernetic

[6] A term originally coined by Tim O'Reilly for a conference to describe the next generation of the Web

[7] See *http://maps.google.com* and *http://earth.google.com* respectively.

[8] *http://www.wikipedia.org*

systems [20]. Contrary to the Cartesian assumptions of classical artificial intelligence, in their study of frog vision, what Maturana and others discovered was that the frog's eye "speaks to the brain in a language already highly organized and interpreted instead of transmitting some more or less accurate copy of the distribution of light upon the receptions" [17].

This discovery caused Maturana to reconceptualize the foundations of cognitive science in terms of autopoiesis: that "living organization is a circular organization which secures the production or maintenance of the components that specify it in such a manner that the product of their functioning is the very same organization that produces them" [20]. First, a frog is autopoietic precisely because its internal metabolism is inside a boundary, frog-skin, that defines its organization as a frog. Second, the components, the organs, are inside the frog-skin and self-reproducing. Yet autopoietic systems are not entirely closed, for the frog's consumption of gadflys and other interactions with the environment are done in lieu of maintaining its own organization as a frog, since eating allows it to bring in energy to maintains its metabolism. A frog adapts its interactions to the environment to maintain autopoiesis [23]. The effect of the world upon any autopoietic system is only an effect insofar as it causes the system to adjust itself in order to maintain its own autopoiesis.

The problem with Maturana's notion of autopoiesis is again the very idea of a unity which implies a cell with a membrane or the skin of a frog. The first condition of autopoiesis, namely that the components of an autopoietic system "through their interactions and transformations continuously regenerate and realize the network of processes (relations) that produced them" suits collective intelligence just fine, as there is no reason a priori why these interactions have to be biological [20]. It is the second part of this definition of autopoiesis that causes us trouble, which is that the components that "constitute it (the machine) as a concrete unity in the space which they (the components) exist by specifying the topological domain of its realization as a network," in other words, "organizational closure" [20]. The problem then is that what autopoiesis explains perfectly is the formation of this topological domain, the crucial fact that cells developed membranes and frogs develop frog-skin that allowed them to separate from their environment. What is lacking is more exploration on precisely how these membranes or boundaries also allow interaction with the outside environment. However, frogs eat gadflies, and grass needs sunlight to grow, and humans use the Web to get directions. How can this bias in favor of the closed system willfully be maintained?

The answer of Maturana and Varela is to introduce the concept of structural coupling to deal with an individual organism's interaction with the environment, a "history of recurrent interactions leading to the structural congruence between two (or more) systems"[21]. Yet instead of two closed systems "perturbing" each other for their mutual autopoiesis, it is easy enough to change perspective to see them as one system maintaining co-evolved autopoiesis. Due to this loophole, even autopoietic systems can become open to the external environment - which after all, are necessary for the system's reproduction - and so open to non-biological organic couplings. The problem with autopoiesis is that precisely as it attempts to get away from a reproductive or genic definition of life that presupposes the individual or the propagation of their genes as the primary feature of life, its definition of organizational closure isolates the system from its environment in a way that prevents the individual from actively assimilating parts of the environment into itself as in the Extended Mind argument. Yet this is incorrect, for closure "has nothing to do with the idea of a materially closed system" since "autonomous systems must be thermodynamically far-from-equilibrium systems, which incessantly exchange matter and energy with their surroundings" [30]. The way out of this dilemma is simple, since closure is used "in its algebraic sense: An operation K exhibits closure in a domain D if every result of its operation yields results within D. Thus, the operation of a system has operational closure if the results of its activity remain within the system itself" [30]. The nub of the problem is that the domain is assumed to be static! Indeed, if the cognitive domain of the autopoietic system can expand to envelop ever more parts that fulfill the two conditions of autopoiesis, then the Extended Mind system can apply to the expansion of autopoietic systems, including heretical bio-social-technological systems.

Despite the biological favoritism of Maturana and Varela, there is nothing inherent in autopoiesis that restricts the components of biology in all possible worlds. The work of Licklider and Engelbart both build from this insight, although they knew nothing of the theory of emergent self-organization, much less autopoiesis as developed by Maturana and Varela. Licklider and Engelbart intuitively grasped that digital computing and representations could easily be part of self-sustaining and intelligent systems. Furthermore, their work led directly to the Internet and the World Wide Web. Instead of aiming to have a machine that is as intelligent as a human individual as in artificial intelligence, Licklider proposed that instead humans and digital computers could couple together closely so that they would become literally symbiotic [19]. Although more work is needed to flesh this case out, it seems there is no inherent contradiction in autopoiesis involving non-biological components. If the individual can be defined via autopoiesis, and to maintain its autopoiesis the individual must increasingly incorporate non-biological components, then the individual is no longer a static, closed system, but an open and dynamic system capable of assimilating and decoupling from various components as it goes in and out of autopoiesis, including digital representations and other biological beings. The obvious objection could be that the biological component is reproducing itself, while the non-biological component is not. Yet is not the reproduction of culture itself reproduction? If so, then humans can be considered not just ways for genes to reproduce, but for our evolving and non-biological technology to reproduce as well.

7 Embodiment Reconsidered

If we now have individuals as open systems that can incorporate non-biological components, do we still have individuals in any useful sense of the term? The main objection to getting rid of the individual would be that the very use of the term embodiment is bound up with that of the biological individual. Before further inspection, the notion of embodiment itself needs to be understood as either simultaneous with or separable from the individual biological body. There does seem something slightly amiss in all the rhetoric of embodiment, as Sheets-Johnstone has pointed out: " the term 'embodied' is a lexical band-aid covering a 350-year-old wound generated and kept suppurating by a schizoid metaphysics" [26]. Everything from a blade of grass to a coffee-cup is embodied in a strictly material sense, and no-one argues otherwise or makes an intellectual programme out of this fact. Embodiment can not just be a synonym for having a physical or material body. What is interesting about embodiment is not the usage of the term as a synonym for the body, but the realization that the context provided by a body can have a causal effect on intelligence. The key word then is "context." N. Katherine Hayles has brought to the forefront that embodiment and the body can actually be spliced into two different concepts. The first, the body, is "always normative

relative to some set of criteria" [12]. In contrast, embodiment is the context that goes along with particular bodies, "enmeshed within the specifics of place, time, physiology and culture...embodiment never coincides exactly with "the body" however that normalized concept is understood. Whereas the body is an idealized form that gestures towards a Platonic reality, embodiment is the specific instantiation generated from the noise of difference" [12]. This is precisely why the notion of embodiment is so difficult for any science to capture, since it is bound up in the very particulars of a given situation that a science or any systematic philosophy must by necessity remove in order to develop any sort of predictive power about future situations and any understanding that applies beyond the here and now. In order to fulfill its role as a science, it is no surprise that cognitive science defined the body by the norm of being bound by the skin. Due to this presupposition, cognitive science has focused more on an a priori "body" than embodiment. Like any fundamentally arbitrary norm, when having to deal with the harsh reality of science, it falls apart. The question returns: If we must construct a body, what kind of body can it be, a body without presuppositions?

As Wheeler relates, when Rod Brooks announced his new paradigm in artificial intelligence based on robotics without representation, in order to distance his positive program for AI from the decades of critique by philosophers, Brooks claimed that at least it wasn't German philosophy [32]. While Wheeler has put together a compelling case that Brooks was in fact doing German philosophy, what we are arguing for is not German philosophy in the vein of Heidegger. To make the case clear, the problem with Heidegger traditionally has not been an emphasis on the biological body. Far from it, since the very Heideggerian notion of the "ready-to-hand" undermines the biological body. Let us look at his paradigmatic example: "the less we just stare at the hammer-thing, and the more we seize hold of it and use it, the more primordial does our relationship to it become, and the more unveiledly is it encountered as that which it is - as equipment. The hammering itself uncovers the specific 'manipulability' of the hammer. The kind of Being which equipment possess - in which it manifests itself in its own right, we call readiness-to-hand" [13]. This readiness-to-hand reveals itself not as abstract knowledge, but as smooth behavior facilitated by the combination of human and hammer. As Wheeler puts it, "the human agent becomes so absorbed in her activity in such a way that she has no self-referential awareness of herself as a subject over and above a world of objects"[32]. At the moment of hammering, given the tight coupling, is it not fair to say that the coupled system of hammer-human is a single system? This is especially true if the hammer is being used in such a way - let's say, to build a house for surviving the cold winter - which is needed for the autopoietic survival of the human agent and his attendant culture, including his hammers in the toolbox. How can even Heidegger himself maintain the biological skin as a crucial boundary?[9] Yet somehow, the very notion of Being is mysteriously tied to the individual human body in Heidegger, and this assumption becomes increasingly uncomfortable, given recent scientific evidence, when refitting cognitive science around on a neo-Heideggerian basis.

To overcome the individual-as-body-in-skin presupposition that is so heavily built into Anglo-American philosophy, what we need is not German philosophy, but French philosophy. French theorists Deleuze and Guattari put forward a concept that can replace the notion of a body: the *assemblage*. In contrast with the individual - even autopoietic - body, Deleuze and Guattari "call an assemblage every constellation of singularities and traits deducted from the flow - selected, organized, stratified - in such a way as to converge artificially and naturally"[7]. Any structural coupling of autopoiesis or instance of the Extended Mind, creates an assemblage. Furthermore, note that this concept is not necessarily disembodied, for the convergence that produces an assemblage can arrive from the "noise of difference," i.e. the context of the world without any abstraction [12]. An assemblage allows us to construct an embodied replacement for the individual body that can keep embodiment while throwing out the individual as an a priori concept. According to Deleuze and Guattari, almost everything in our everyday ontology is an assemblage. In fact, the question then becomes what "bottom-outs" an assemblage, and how to determine if an assemblage exists at a given moment. One further notion brought up by Deleuze and Guattari is that of the *body without organs* that is "under way the moment the body has enough of organs and wants to slough them off, or loses them." The "body-without-organs" allows us to conceptualize bodies as not necessarily biological (i.e. built of organs). More importantly, the body without organs captures the dynamic activity of an assemblage that makes it cast off its previous couplings, and create new ones dynamically in response to its situation. This is the opposite of any statically construed normative body, for the body-without-organs "is not at all the opposite of the organs. The organs are not its enemies. The enemy is the organism" [7]. Let us correct them here: the enemy is not the organism, but the organism as a reified a priori individual.

It seems we have painted ourselves into a corner: If all bodies are collective autopoietic assemblages, then why any assemblages to begin with? If we are throwing away any static body-bounded-by-the-skin, why do the bodies recompose into different collective assemblages, some autopoietic, others not? The answer is in our definition of the body; the source of every conception of the body is inherently normative. Norms do not drop out of the sky as if given to us by the angels; the only scientific story we can tell about norms is evolutionary. As Dennett puts it, all norms must eventually ground out in evolution, although the jury still seems out on whether or not evolution selects genes, individuals, or groups of individuals sharing traits [29]. In a Heideggerian note, the formation of assemblages happens in response to the encountering of problems thrown our way by the world, and our attempt to maintain the autopoiesis of these assemblages as they are faced by these problems, ranging from fleeing sabre-tooth tigers to collectively avoiding extinction of the species. Success in these problems is measured in evolutionary terms, whether or not the assemblage can survive and maintain autopoiesis. As the problems change, so will the assemblages. The assemblage of cells known as the biological human body incorporated the assemblage known as the skin as a solution to problems of heat regulation, evaporation adaption, self-defense, and other problems encountered by cells trying to maintain their autopoiesis. Furthermore, this evolutionary story can be harnessed to explain the emergence of collective intelligence in the forms of the Web. Problems today, ranging from mapping the genome to prevent disease to the co-ordination of production and consumption in a globalized market, are far beyond the knowledge and representations easily accessible without the heavy-duty cognitive scaffolding of the Web. The development of the collective intelligence is the only way to harness the fact that "no one knows everything, everyone knows something, all knowledge resides in humanity" [18]. We can detect the formation of new assemblages, and the representations they utilize and incorporate, by paying attention to the problems that threaten the previously stable assemblages.

[9] Although, we might add that Heidegger does make the human body to be "wonder tissue," by regulating hammers and whatnot to "equipment" and denying Dasein to all but humans. Further explication of this would be illuminating, but is beyond the scope of this paper.

8 Conclusions

Conservatively, what we have argued is two-fold. First, that the notion of representations championed by Brian Cantwell Smith can be built on top of neo-Heideggerian notions of embodiment, and this allows phenomena such as the Web to be brought under consideration and explained as sources of intelligence. Second, and more radically, the assertion that "online intelligence is generated through complex causal interactions in an extended brain-body-environment system" can be pressed in such a way that we can philosophically "come out on the other side" and end up in a world that allows collective intelligence built on top of distributed representations. This allows philosophy to escape from the confines of an overly restricted embodiment that is restricted to the biological body, and so "an increased level of biological sensitivity" in cognitive science should be complemented by an equal sensitivity to the non-biological aspects of intelligence. Humans and animals are systems embedded in a non-biological culture - and that matters for cognitive science. Lastly, while we would not argue against the priority of online intelligence per se, we would hope that it does not miss out the fact that increasingly online intelligence is incorporating the heavy use of representations and other aspects of what has traditionally been thought of as "offline" intelligence. Think about the difference between scavenging for nuts and berries and navigating hyperlinks on the Web to discover a map to the grocery store, for as both Deleuze and McLuhan would note, the Web is the return of the information gathering nomad. This also undermines any methodological insistence on a dynamical system analysis, since dynamical systems have shown trouble in handling anything that appears to be a representation and while they are useful in modelling, they are trapped by their own dependence on initial parameters that may or may not be scientifically illuminating [25].

In a more radical direction, we have questioned the biological body as the useful level of analysis for cognitive science, and so a simplistic version of the neo-Heideggerian embodiment programme as pushed by the work on robotics by Brooks [3]. The body is not given, but is created dynamically as a collective assemblage justified in terms of the problem at hand, where success at the task at hand is grounded out in the normativity of evolution. This has certain resonances with work in continental philosophy, in particular Deleuze and Guattari. Defining intelligence in terms of a fully autonomous agent is not even an accurate portrayal of human intelligence, but a certain conception of the individual human subject, "a certain conception of the that may have applied, at best, to that faction of humanity who had the wealth, power, and leisure to conceptualize themselves as autonomous beings exercising their will through individual agency and choice"[12]. By jettisoning this conception, and maintaining the commitment to a certain necessary degree of embodiment as given by a rehabilitated neo-Heideggerian programme, cognitive science can do justice to complex phenomenon such as the advent of the Web and increasing recognition of collective intelligence. Levy notes that cognitive science "has been limited to human intelligence in general, independent of time, place, or culture, while while intelligence has always been artificial, outfitted with signs and technologies, in the process of becoming, collective"[18]. The vast technological changes humanity has engendered across the world are now reshaping the boundaries of human bodies, and so the cognitive world and the domain of cognitive science. This has been a process that has been ongoing since the dawn of humanity, but only now due to the incredible rate of technological progress, as exemplified by the growth of collective intelligence on the Web, does it become self-evident.

REFERENCES

[1] Tim Berners-Lee. IETF RFC 1630 Universal Resource Identifier (URI), 1994. http://www.ietf.org/rfc/rfc1630.txt.

[2] Tim Berners-Lee, Roy Fielding, and Larry Masinter. IETF RFC 3986 Uniform Resource Identifier (URI): Generic Syntax, January 2005. http://www.ietf.org/rfc/rfc3986.txt.

[3] Rodney Brooks, 'Intelligence without representation', *Artificial Intelligence*, **47**(1-3), 139–159, (January 1991).

[4] K. Chen, L. Fine, and B. Huberman, 'Eliminating public knowledge biases in information-aggregation mechanisms', *Management Science*, **50**, 983–994, (2003).

[5] Andy Clark, *Being There: Putting Brain, Body, and World Together Again*, MIT Press, Cambridge, MA, 1997.

[6] Andy Clark and Dave Chalmers, 'The extended mind', *Analysis*, **58**(1), (1998).

[7] Gilles Deleuze and Felix Guattari, *A Thousand Plateaus*, University of Minnesota Press, Minneapolis, Minnesota, 1987.

[8] R. Dunbar, *The Human Story*, Faber and Faber, London, United Kingdom, 2004.

[9] V. Gallese and A. Goldman, 'Mirror neurons and the simulation theory of mind', *Trends in Cognitive Science*, **12**, 493–501, (1998).

[10] Harry Halpin, 'Representationalism: The hard problem for artificial life', in *Proceedings of Artificial Life X*, (2006).

[11] John Haugeland, 'Analog and analog', in *Mind, Brain, and Function*, Harvester Press, New York, NY, (1981).

[12] N. K. Hayles, *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature and Informatics*, University of Chicago Press, Chicago, Illinois, 1999.

[13] Martin Heidegger, *Being and Time*, Blackwwell Publishing, London, United Kingdom, 1927. translated by John Macquarrie and Edward Robinson (1962).

[14] F. Heylighen, M. Heath, and F. Van Overwalle, 'The emergence of distibution cognition: a conceptual framework', in *Proceedings of Collective Intentionality IV*, (2004).

[15] Peter Hoshka, 'CSCW research at GMD-FIT: from basic groupware to the social web', *ACM SIGGROUP Bulletin*, **19**(2), 5–9, (1998).

[16] E. Hutchins, *Cognition in the Wild*, MIT Press, Cambridge, Massachusetts, 1995.

[17] J. Lettvin, H. Maturana, W. McCulloch, and W. Pitts, 'What the frog's eye tells the frog's brian', *Proceedings of the Institute Radio Engineers*, **47**(11), 1940–1951, (1959).

[18] Pierre Levy, *Collective Intelligence: Mankind's Emerging World in Cyberspace*, Plenum Press, New York City, New York, 1994.

[19] J. Licklider, 'Man-computer symbiosis', *IRE Transactions on Human Factors in Electronics*, **1**, 4–11, (1960).

[20] Humberto Maturana and Francisco Varela, *Autopoiesis and Cognition: The Realization of the Living*, D. Reidel Publishing, Dordecht, 1973.

[21] Humberto Maturana and Francisco Varela, *The Tree of Knowledge*, Shambhala Press, Boston, Massachusetts, 1987.

[22] A. Newell, 'Physical symbol systems', *Cognitive Science*, (4), 135–183, (1980).

[23] Ezequiel Paolo, 'Autopoiesis, adaptivity, teleology, agency', *Phenomenology and the Cognitive Sciences*, **4**(24), 429–452, (2005).

[24] A. Pentland, 'On the collective nature of human intelligence', *Adaptive Behavior*, **15**(2), 189–198, (2007).

[25] Luis Rocha and Wim Hordijk, 'Material representations:from the genetic code to the evolution of cellular automata', *Artificial Life*, (11), 189–214, (2005).

[26] M. Sheets-Johnstone, 'Emotion and movement', in *Reclaiming Cognition*, eds., R. Nunez and W. Freeman, Imprint Academic, (1999).

[27] Herbert A. Simon, *The Sciences of the Artificial*, MIT Press, Cambridge, Massachusetts, first edn., 1969.

[28] Brian Cantwell Smith, *The Origin of Objects*, MIT Press, Cambridge, MA, 1995.

[29] K. Stotz and P. Griffiths, 'Genes: Philosophical analyses put to the test', *History and Philosophy in the Life Sciences*, **26**, 5–28, (2004).

[30] Evan Thompson, *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*, MIT Press, Cambridge, Massachusetts, 2007.

[31] M. Tomasello and H. Rakoczy, 'What makes human cognition unique? from individual to shared to collective individuality', *Mind and Language*, **18**(2), 121–147, (2003).

[32] Mike Wheeler, *Reconstructing the Cognitive World: The Next Step*, MIT Press, Cambridge, Massachusetts, United States, 2005.

Constructivism in AI: Prospects, Progress and Challenges

Frank Guerin[1]

Abstract. This position paper argues the case for the application of constructivist theories to Artificial Intelligence, with particular emphasis on Piaget's theory. The idea of building an artificial baby is an old one in Artificial Intelligence, yet it is difficult to execute because little is known about the information processing mechanisms which babies use to learn. That which is known comes from non computing disciplines and has not been exploited very much in Artificial Intelligence. Part of the difficulty is that many AI researchers do not know enough about these other disciplines; another difficulty is that where AI researchers do know some of the theories from other disciplines, many do not see their value. This paper tries to make a case for the value of Piaget's theory in particular.

1 INTRODUCTION

This paper sets forth a research agenda for the application of constructivism to Artificial Intelligence. We take the view that there is a mine of useful ideas in constructivist theories, particularly Piaget's[2], that have yet to be fully explored by AI researchers.

Section 2 gives a brief description of what is meant by a "constructivist approach to AI", and contrasts it with non-constructivist approaches. It uses the behaviour of "groping with a stick to retrieve an object" as an example task, to contrast the two approaches. It then gives a fairly detailed account of how this behaviour could be acquired by an AI system which follows Piaget's theory. This account serves to give the reader a detailed picture of what is entailed by the constructivist approach. Section 3 makes some arguments to justify why we believe the constructivist approach to be worthy of investigation, and tackles some common objections. Section 4 reviews some of the existing work in constructivism in AI, to see what has been achieved and what remains to be done. Section 5 briefly outlines some of the major challenges to be tackled in following the constructivist research programme advocated here. Section 6 concludes.

1.1 Controversies about Piaget's Theory

There is considerable controversy over the claims Piaget has made, particularly when it comes to what knowledge is innate or learned. Some results in the 1990s purported to show a great deal of innate knowledge, far beyond what Piaget had proposed. However, researchers in the constructivist camp followed up with studies of their own and drew different conclusions. The controversy is a lively one, with claims of improper experimental procedures, overinterpretation of results, failure to replicate reported results, etc. This paper will not delve into the controversy, interested readers could consult Cohen and Cashon [5] and Haith [8] as a starting point. This controversy looks unlikely to be settled in the near future; in the meantime, we must concede that it is not certain that constructivists' theories of human learning are an accurate account of what humans actually do, but we take the position that it is worth trying it out for AI. Apart from possibly helping us to build more intelligent programs, an AI investigation might also shed light on whether or not constructivism is a viable theory to explain human learning.

[1] University of Aberdeen, Scotland, e-mail: f.guerin@abdn.ac.uk

[2] This paper draws heavily on Piaget's account of infancy, as described in his trilogy of books [10, 12, 11]. Where specific references are not given, we are referring to this trilogy, and usually to the first two books [10, 12].

2 CONSTRUCTIVISM IN AI

This section explains what we mean by constructivism in AI. We use an example behaviour from Piaget's observations, and explain how a constructivist approach to AI might go about implementing it, mostly following Piaget's theory of constructivism.

Piaget describes the "behaviour of the stick": the infant seeks to take possession of an object which is located out of arm's reach; the infant uses a stick as a tool to draw the object into the range of his arms, and then takes possession of it. To keep things simple we can restrict our attention to the case where the infant is provided with the stick by an adult. In this case the infant typically becomes capable of the behaviour of the stick roughly between 12 and 18 months (Piaget's fifth sensorimotor substage). The case where the infant has not been given the stick, and must think of the idea to use it himself, is more advanced and belongs to the subsequent substage. Note that being provided with the stick, and being shown an example of the required behaviour, does not at all make the task trivial for the infant. An infant who has not yet achieved the behaviour of the stick may attempt to copy the adult behaviour, but can only do so coarsely and will merely strike the desired object, and be unable to draw it closer; the infant at this stage will limit himself to repeatedly striking the object in the same way, and will not attempt to modify the way that the object is being struck. This primitive behaviour may persist for several weeks before the infant properly begins to grope and thereby acquires the behaviour of the stick. By groping we mean that the infant modifies the way of striking, and so identifies the type of striking that brings it closer.

We will now contrast two possible approaches to implementing the behaviour of the stick in an AI system: firstly a non-constructivist approach which makes use of prior knowledge, and secondly a constructivist approach where the infant must construct the relevant knowledge. For the non-constructivist approach we could employ a reinforcement learning algorithm. The artificial infant should get a small reward whenever the stick is brought closer to the object; this could allow the infant to quickly learn to bring the stick in contact with the object. Upon contacting the object with the stick, the infant should get a large positive reward whenever the object is moved closer to the infant, and a negative reward whenever the object is pushed further away. Through random groping the infant will find the actions which draw the object closer. This is all doable with present

techniques, and will lead to the artificial infant retrieving objects with reasonable efficiency. The main "cheat" being exploited here is that we have told the infant (via rewards) that it needs to bring the stick in contact with the object, and that it needs to take certain actions which make the object come progressively closer; we have also provided the infant with information about when the object is coming closer, so it does not need to worry about knowing how far away the object is. All of this amounts to giving the artificial infant prior knowledge about the world, and specific knowledge about how to achieve a particular goal.

The constructivist approach would try to avoid giving the infant any prior knowledge beyond that which is absolutely necessary to bootstrap the learning process and allow the infant to learn in reasonable time[3]. Apart from this minimal innate knowledge, the constructivist approach aims to allow the infant to create the required knowledge for itself. For example, innate knowledge might include the ability to grab that which touches the hand, the ability to suck that which touches the mouth, the ability to make random arm movements, etc. By bootstrapping from these initial abilities, the infant must learn how to suck the thumb, how to grab and suck world objects, and how to interact with world objects in more complex ways. Through this interaction the infant must somehow learn higher level knowledge about the world, gaining knowledge of space and objects and how to manipulate them. This would eventually lead to the behaviour of the stick.

The path along which an infant develops the required knowledge has not yet been determined by psychology (i.e. what specific episodes lead to the acquisition of what specific knowledge). Piaget's theory gives a sketchy overview, but many details remain to be fleshed out. Coming up with *(1) a plausible path of development* is the first challenge for a constructivist approach to implementing the behaviour of the stick in an AI system. A speculative and somewhat sketchy account of such a path is given in the next subsection. Given such a path the next step is to *(2) computationally model the information processing*, i.e. code learning algorithms which can profit from the experience specified by episodes in the path, in order to learn the required knowledge. How this can be programmed is not at all clear with present day AI techniques. This is the second challenge for constructivism in AI. Section 4 reviews some related work which has developed algorithms for some types of constructivist learning.

2.1 A Possible Development Path

This section gives a highly speculative account of how an infant (real or artificial) might develop the behaviour of the stick, starting out with only some very basic innate abilities. It follows Piaget's theory very closely, with some minor variations to explain (roughly) how a computer system might learn it. It is speculative because it has not been tested either with real infants or AI simulations. This is the case with Piaget's books on infancy; he proposed a theory which accounted for what he observed in his own three infants, but he did not attempt to falsify it through larger studies. It must be conceded that the description is also sketchy in places, and many details would need to be worked out to create a computer implementation. The purpose of presenting this fairly detailed account here is to give the reader a clearer idea of what we mean by a constructivist approach in AI, and to highlight some of the important aspects of the constructivist point of view. Whether this account is correct or not, some description along similar lines could be proposed and tried out.

The acquisition of thumbsucking seems not difficult to account for: when the infant notes that his hand has come into contact with his mouth, he can remember the movement which brought it there from some near location x, and so repeat it in future, if he finds his arm at x (he can recognise this location x by proprioception). In a similar way he can find the action which brings his hand from some more distant location to x. Extending this idea through all the space in which the arm moves, the infant can find actions which bring his hand to his mouth from any initial arm location. With this ability acquired, one can see how he can learn to grab a touched object and take it to his mouth. One can also imagine how this can be extended so that he learns to focus on some object with his eyes, and then to move his hand to the centre of his vision to grab it, and then suck it. This is a significant development; it leads the infant to observe new interesting phenomena; new unexpected things will happen, because he is now interacting with the external world (as opposed to simply sucking a part of his own body). For example, when reaching with his hand to grab a hanging object, he may accidentally knock against it, producing a swinging motion. This interests the infant and he tries to rediscover the action that caused it. In this way he discovers a motion which can effectively swing the hanging item; he refines this movement to develop an effective "striking" action, to augment the swinging of the item. This strike action, being new, is now tried out on every object the infant encounters. The infant is generally eager to try our any new actions as much as possible. A number of new actions are developed at this stage, but we can restrict our attention to striking.

The effort to grab objects in the external world also inevitably leads the infant to confront problematic situations; for example, situations where the desired *object* is visible, but some *obstacle* blocks the infant's hand from grabbing it. The solution to this problem is to simply push aside the obstacle, however this is not at all obvious to the infant at this stage. There are scenarios where the infant could use his "striking" action to knock aside the obstacle, yet he does not. This reveals a lot about the infant's world model at this stage. Each action in the infant's repertoire is isolated; there is no coordination between them, and hence no knowledge of their relationships. Thus the infant's world model consists of a series of fragmentary spaces: one to describe the effect of each individual action. He knows well that striking the obstacle will move it in a certain direction. Why is it that he cannot strike and knock aside the obstacle in order to clear the path and reach for the desired object? It is because if he thinks of striking aside the obstacle, then he is considering this striking action along with its fragment of space (by this we mean that he has knowledge of the postcondition of the action: that the struck object will move away from him in space); the obstacle is represented in this fragment, but the desired object is not. So although he could foresee the obstacle moving, that "mental image" does not include the object, so he cannot foresee that the object would become accessible. To foresee this he would need to represent both the object and obstacle in his mental image. That presumes a more sophisticated representation of space which has connected up the two fragments. In fact this "connecting up" of space will happen through the coordination we are about to describe. By chance, on one of the failed attempts to grab, and upon clashing with the obstacle, the infant may decide to strike the obstacle, without foreknowledge of what this will achieve.[4] Having done this, the infant has not forgotten his recent goal to grab the object,

[3] We do not want to make the learning problem so difficult that the time required is inordinate; e.g. it would take too long if we had a reinforcement learner that only got reward when the object was retrieved.

[4] This departs from Piaget's account which describes it as an intentional act to "negate" the inclusion of the offending obstacle in the grabbing scenario.

and now sees that the path is clear, and so grabs it. This coordination is remembered, and used thereafter. The infant also learns to discriminate the visual appearance of a scene where an obstacle to grabbing is present from one where an obstacle is absent; using this knowledge the infant is able to employ the intermediate striking action in those cases where it is required, without having to first bump against the obstacle on a failed grab. This striking action, to remove an obstacle, gets used repeatedly, and becomes refined, gradually leading to a displacing action which can pull an offending obstacle aside. With these acquisitions the infant is now able to pay attention to the relationship among any two objects which are seen, and especially to notice when one is "in front" of the other, or could be moved "in front" of the other. The acquisition of this displacing action means that he has an action in his repertoire which can change the relationship among two objects. This action is of a higher order than those previously in his repertoire; in Piaget's language: it belongs to the next stage. Having acquired this behaviour, it can again be extended: the infant can learn to bring one object closer to another, when that causes an interesting interaction between them which he would like to repeat.

Let us now look at the behaviour of the stick and see how far we are from achieving it. By now the infant could strike an object with the stick if given the stick and shown an example of the behaviour; however, he could not draw the object towards him. His limitation at this stage is that when he strikes the object, and it moves a little, he is unable to interpret this movement (the movement is meaningless to him and he is effectively blind to it). A more advanced infant would understand that if it moves a little because of the stick, then the stick has the power to displace it in space, and further groping might well find actions that move it on a trajectory towards the infant. In order to be able to use the stick as a tool to bring a desired object closer, the infant will need to know the trajectory on which the desired object should travel in order to come within range of his arms. This presumes a knowledge of where the object is currently, and the location to which the infant would like to bring the object. Surprisingly, the infant at this stage (8-12 months roughly) does not actually know where an out of reach object is, even if it is seen. Knowledge of location in space will need to be constructed. Understanding this is crucial to understanding the constructivist's hypothesis about the infant's view of the world (or knowledge of space in this instance), and its development.

Piaget describes the child's increasing knowledge of space as akin to an expanding sphere. By investigating the space accessible to his arms he comes to have a practical knowledge of this "near space"; practical in that he knows how to grab something he sees in this space, but he does not understand how different locations in this space are linked. At this stage looking at objects in far space (out of reach) is like looking at images painted on the inner surface of a sphere, which are mostly static, but sometimes move in mysterious ways. An object which moves from near space to far space does not only change its position for the young infant, to the infant it is as though it changes its nature. The infant at this stage only reaches for objects in near space and does not bother with those out of reach. With development he begins to see these objects as constant and begins attempting to reach for objects which are just out of range. When he begins to pay attention to relationships among two objects in near space (a development which could be reached by the path described above), he then begins to notice the parallax caused by motion of his head. Knowing the relative positions of two objects A and B in near space, and the parallax caused by head movements, he can make the conjecture that there is a correlation between parallax and relative position. When he then sees the parallax between a near object C and one which is just out of reach D, he can reason by analogy that the far object D is probably behind the near object (i.e. because the parallax observed between C and D is the the same as that between A and B therefore the spatial relationship is probably also the same). Thus he begins to locate the far object D in space. This development can then be extended to notice that a further object must be located behind D, and so on. In this way he can begin to locate the position of all the objects in a room relative to one another. However limitations will still exist for very far objects. The expansion of the child's sphere happens over years, with young children wondering why the moon follows them, before reasoning that it is merely the parallax motion between it and the trees.

It may be difficult to imagine the infant's primitive view of the world, given the things we take for granted from our perceptual experience. It is perhaps easier to understand the development described above by looking at how adults can also extend their knowledge of space by expanding the sphere in which distances can be reckoned. Primitive societies tend to view the stars as points on the inside of a hemisphere, sometimes the sun and moon are also thought to be the same distance from the Earth. Aristarchus of Samos was able to reckon that the Sun was much further away than the moon by considering the angle between the sun and moon when the moon is half-lit (imagine that the sun is a lantern, the earth an apple and the moon an orange, and then consider the angle made by lantern-apple-orange, when the orange is half lit from the apple's perspective) . Thus he was reasoning by analogy with his knowledge of illuminated spherical objects in the near space where distances were known to him. He hypothesised that the same laws held for far objects. In more recent times telescopes have allowed parallax to be used to estimate the distances to stars, by measuring angles from different orbital positions of the earth. This is an example of constructivist learning; it the application of known facts (diameter of Earth's orbit, geometry) together with new observations (angles subtended by stars in Winter and Summer) to construct new knowledge (distances to near stars).

Returning to the infant and the stick, we can now see how the infant could come to know the location of an object which is out of reach. He can reason relative to successive landmarks, starting from some location in near space, and going to the desired object, each being behind the other. When the desired object is struck, the infant can note any movement it makes relative to nearby landmarks (even elements of a pattern/texture on the floor could serve as landmarks). This relative movement is now salient; in contrast to the infant at the earlier stage, any relative movement between two objects is now interesting, because the infant knows an action which can do that, and has played with changing relative positions. The infant can understand that if the stick can cause a small relative movement, then a movement can probably be found to bring it closer relative to some nearby landmark; this could then be repeated to bring the object in front of successive landmarks and thus into near space.

An objection may be raised here to say that the infant does not need landmarks to precisely locate objects in space, because he could gauge the distance to which his eyes focus and the direction in which they point. This would require that the infant had carefully calibrated his eyefocus to match it against different distances, and had noticed the analogy between changing focus and changing parallax (for example), to conjecture a relationship between changing focus and distance. This is a possibility, though it seems unlikely; it would be difficult even for an adult to reliably tell if an object has moved closer or further away when it makes a small movement in some random direction in a featureless space (contrast with the object moving on

a patterned carpet). Similar arguments and counterarguments can be made for attempting to locate an object in space based on its apparent size.

This then is what we mean by a constructivist approach to learning the behaviour of the stick. In particular constructing the knowledge of space requires quite a sophisticated type of reasoning by analogy. Constructivism proposes that the kind of learning machinery which can do such analogical reasoning is innate in the infant. Thus a constructivist approach to AI escapes the burden of coding a great deal of innate knowledge, but takes on the burden of coding a very sophisticated learning algorithm. The scientific objective of constructivism in AI is to discover what type of information processing mechanism could implement the constructivist theories of Piaget and others.

3 PROSPECTS: The Case for Constructivism

Having introduced the basic ideas of what we mean by a constructivist approach in AI, we now make a case for why it may be a promising approach. We will support the position by considering some arguments for and against. In comparing the constructivist and non-constructivist approaches to learning the stick above, one may be struck by the extremely poor level of initial knowledge of the constructivist system, lacking even the ability to tell where visible objects are located in space. It is reasonable to wonder if there is anything to be gained by handicapping the system so severely. The counterargument: the point of the proposed approach above is really to advance our scientific knowledge of constructivist learning; our ultimate goal is not to achieve the behaviour of the stick, else we would code the required competence directly. By handicapping the system in terms of prior knowledge in this task we are forced to come up with a system which can gather the required knowledge through its interactions with the environment. Thus we are hoping that by making the task specification of the problem roughly similar to that which faces the infant, we may be able to come up with an algorithm which has some similarity (at a high level perhaps) to what the infant uses to solve the problem; hence it should be extensible and go beyond the specific tasks which it was trained on.

Our interest in constructivist learning is not purely to advance our knowledge of cognitive science: we expect that constructivist learning mechanisms will prove useful in practical AI systems; this position will now be supported. We expect that endowing AI systems with constructivist learning mechanisms will bring two benefits: (1) it is likely to be a good approach to the commonsense knowledge problem (i.e. let the program learn itself with constructivist mechanisms); (2) it is likely to be a good approach to allowing systems to generalise from what they know and so learn how to cope with new situations (in fact it treats these issues as central).

Point (1) claims that it will be easier to build a constructivist learner and let it acquire the commonsense knowledge of a three year-old (for example) through interactions with the world, than it will be to code that knowledge directly. A counterargument could be that perhaps constructivist learning is just some idiosyncratic way in which the child seems to learn. Given that we know a lot more about objects and space, perhaps we could code our advanced knowledge directly into our programs, and then they would have no need to perform constructivist learning. This counterargument has some force, because if we consider constrained tasks, then constructivist approaches are certainly not the best. For example, industrial robots can perform highly skilled operations which require some knowledge of space and objects, and these robots do not have to go through a long apprenticeship, as the child does. Industrial systems often achieve tasks which humans could also do, but in a very different way. Their representation of space, for example, is clearly very different to that which would be built by constructivism, yet it works well for the tasks these industrial systems have to do. However, experience has shown that, for general knowledge, it does seem to be exceedingly difficult to code all the commonsense knowledge of even a child into a computer. Existing approaches which have been successful in constrained domains have not proved to be extensible to general knowledge. It is worth trying alternative approaches.

It is not clear if constructivism is the best approach to learning about the world (and space for example), but it is one way that works (in the human). One of its promising aspects is that it does seem to give a plausible account for some of the human abilities which current AI systems lack, and some of the idiosyncrasies of human reasoning. Humans often make incorrect analogies, as can be seen from the history of science and pre-scientific notions. Consider the following gem from Piaget's investigations; it is a child's response to a question about why a helium balloon goes up: "Because there's a gas inside, when there's a lot of gas it's heavy, it's very strong and then it flies." Though the constructivist mechanism can sometimes lead to wrong conclusions, it is perhaps more surprising how often it leads to useful conclusions. The child quoted clearly has some intuitive notion of inner force or strength, and though this is for the moment confused with weight, it is nevertheless a concept which will lead the child to correct deductions in many cases. It is precisely this type of intuitive concept which present AI systems are sorely lacking. The constructivist approach we advocate forces us from the outset to find representations for knowledge which are extensible, and which facilitate analogical reasoning (otherwise the development path we attempt to model will be unachievable), thus the hope is that we will come up with representations for intuitive concepts such as this. In contrast, in a non-constructivist approach we would focus on achieving a specific competence, rather than modelling a development path. Then we code in knowledge directly into the system, and the danger is that we do it wrongly; i.e. we represent things in a way which is good for that particular competence, but not very useful for performing general tasks, for drawing conclusions or for extensibility.

As for the usefulness of the knowledge coded, it is interesting to note that in the Piagetian account of the construction of space, it is constructed by assembling fragmentary spaces, where each is a fragment of knowledge about a known action. The upshot of this is that when the space is constructed, and when the infant sees a distant object, he immediately knows the actions which could manipulate it relative to some nearby landmarks. (He may not yet know the behaviour of the stick, but he knows that if he were within reach of the object he could manipulate its position relative to those landmarks.) He will know an action to move it closer relative to some nearby landmark (that is the action of the hand drawing an object closer relative to a landmark). In fact the object's position is effectively represented in terms of a series of actions which could bring it to the infant, or alternatively, the moves the infant needs to perform to get there. Piaget cites a pertinent quote from Poincaré to support his theory about the conception of space: "to localise an object merely means to imagine the movements which must be made in order to reach it". More generally, to perceive a scene, according to Piaget's theory, is to simultaneously be aware of all the different actions that could be applied there. This is because the construction of the actual objects perceived is performed by recruiting a host of low level fragments of sensorimotor knowledge (i.e. construction of a perceived object from the image sensed). In looking at a distant building, all that is sensed are a few points of light on the retina, but the mind elaborates this so

that a three dimensional building with inside, outside, rear, etc. is perceived. This elaborate 3-D structure is constructed in the mind by making use of our past experiences when we walked inside buildings, upstairs, around the back, etc. and indeed many more general experiences which have discerned the solidity of materials, dimensions, positions in space, etc. [10, see especially p. 189-190]. This is what Piaget means by the construction of reality. It explains how human visual perceptual experiences could be quite different from standard approaches to computer vision (see also [14]). Computer vision seeks to reconstruct the 3-D structure that is being viewed by 2-D cameras, and to recognise objects there. However the human version considers all the past actions that were applied on similar shaped surfaces, and it is largely in terms of these actions that the 3-D representation of the scene is reconstructed. Therefore, when the human seeks to manipulate what is seen, to achieve some goal, the objects and actions that could achieve that can immediately spring to the forefront for consideration. In this way the constructivist approach to acquiring world knowledge holds the promise of answering Sloman's concerns about "affordances", and the *understanding* of surface structure [14].

The idea of advanced knowledge being built on top of simpler well tested knowledge (i.e. sensorimotor schemas at the lowest level) applies to all constructivist acquisitions. The contrast between a constructivist knowledge representation in AI, and a classical AI representation to solve the same problem is similar to the contrast between a student rote learning how to perform a particular mathematical operation, or a student understanding the operation as a new combination of operations he has already mastered. In the case where the student can construct the new operation as a function of known simpler operations, the knowledge is much more useful, and can be adjusted and applied in diverse situations; in contrast, the rote learnt procedure would be rigid and only useful in a narrow set of situations. Piaget's statement that "to understand is to invent" is pertinent here. To take this argument back to AI, one could criticise attempts to code knowledge such as naive physics into an AI system, because a system with such knowledge would not understand how to apply it, when it is not grounded in its own more primitive sensorimotor knowledge.

The argument in support of point (2) above continues the idea of analogy and extensibility introduced to support point (1). In the case of (2), as opposed to the commonsense knowledge case, we are considering scenarios where the knowledge required may be unknown to the system builder at design time, so having the designer coding the knowledge is not an option. Constructivism makes no distinction between learning commonsense knowledge (1) or generalising from known facts to learn to cope in a new situation (2). They both require the learner to conjecture a new theory and to test it, and refine it as necessary. The same mechanism is used in both cases. Thus the constructivists' claim is that the machinery for generalising and analogy forming which is required to form a solution to a difficult new problem is one and the same as that required to learn the basic world knowledge which children learn. The claim is that infants are doing scientific discovery all the time. Examples from the study of scientific discoveries provide support for the constructivist claim; analogy with previous situations seems to be the key to conjecturing new models of how the world works. It would be difficult to provide any alternative account, because a human can do nothing other than conjecture based on things already known. To conclude this argument: It is impossible to say if constructivism is the best way to learn, but there do not seem to be any viable alternative accounts of how this type of knowledge acquisition could proceed.

3.1 A Common Learning Mechanism?

An objection is sometimes raised over the constructivist claim that there is a common learning mechanism which is being used by the adult to discover new knowledge, and by the infant to learn basic world knowledge. The objection is that surely the adult has a superior ability when it comes to adding knowledge; surely the functions the adult uses are different to those of the child. The adult has a logical framework in which hypotheses can be entertained and dismissed, and contradictions can be noted. The child is often comfortable with giving two somewhat contradictory explanations for a phenomenon. For example: large boats float because they are very heavy and strong and can push down the water, while stones sink because they are heavy and strong and can push apart the water to get to the bottom. The response to this objection is that the adult has more learning tricks in addition to the core learning mechanism which is common with the child's; these extra functions of the adult were in fact built by the core learning mechanism.

The same phenomenon can be seen in evolution. The basic mechanisms of evolution are very simple: replication, mutation and natural selection. But as evolution progressed it evolved fancier tricks. For example, there are two types of genes: structural genes and regulatory genes [13, Ch. 20]. Structural genes code for "building block" proteins for example, while regulatory genes control the expression of other genes. Regulatory genes can act like switches, controlling structures that appear on the body. The evolution of genetic switches makes subsequent evolution easier. It can be thought of like a computer program becoming modular, where modules can be called with parameters; it makes modification easier. Sexual reproduction is another example of a fancy trick by which evolution has been able to accelerate its own learning. These tricks are not present in early life, and do not need to be; the basic mechanisms of evolution are enough, the other tricks will appear through evolution. In the same way, constructivism claims that many of the fancy learning tricks used by adults do not need to be present in the innate learning mechanism.

3.2 Why Copy Infants?

An objection may be raised over the choice to copy acquisitions in infancy, as opposed to acquisitions in childhood or adulthood. Piaget's theory hypothesises that the same learning mechanism is in use at all ages; given that we are really interested in modelling the learning mechanism, rather than a particular acquisition, we could well model some learning episode in an adult. The UK Computing Research Committee's Grand Challenge 5 meeting reports that "it was argued that newborn infants are much harder to study since most of what they do is very inscrutable" [5]. It is true that infants are difficult to scrutinise because you cannot ask them what they are thinking, however children and adults could be viewed as equally inscrutable because they have subconscious processes which are not open for introspection, and sometimes they can give very inaccurate accounts from introspection. An argument can be made that infants may be more scrutable, because at a young age they are incapable of simulating a possible course of action in their heads, and must try it out by groping in the real world to see what happens. Furthermore, at this stage they may use their body to physically represent a situation they are dealing with. Piaget recounts an observation of one of his daughters who is attempting to put a chain into a matchbox which is open just to a 3mm slit. Not succeeding in making the chain enter, she seems to represent the slit with her mouth, opening her mouth

[5] http://www.cs.bham.ac.uk/research/projects/cogaff/gc/

wider and wider, and thus discovering a solution to the problem, i.e. to enlarge the opening, which she promptly does. After 18 months the infant begins to solve these types of problems covertly, simulating them inside his head. From this age the infant certainly seems to become more inscrutable, when many ideas are experimented with internally, and the infant suddenly comes up with a solution to a problem, seemingly out of nowhere.

More than the above argument though, the main argument in support of modelling acquisitions in infancy is that it should be easier, because knowledge is being built on a smaller set of existing knowledge. The acquisition of new knowledge requires finding analogies with existing knowledge, so we need to trace out exactly how each new acquisition is built on previous knowledge. This can be very complicated; especially when there are so many possible pieces of existing knowledge that a new piece of knowledge could relate to. Our task (in pursuing the constructivist approach to AI) is to come up with the mechanism which can add the new knowledge. We need a precise specification of this task in order to write an algorithm to do it. The task is: given a certain starting state of knowledge, and a new experience which goes beyond (or is not consistent with) existing knowledge, add some new knowledge which accounts for (is consistent with) that experience. One of our challenges is that it is difficult to know the starting state of knowledge, and having an accurate description of this state is crucial, because if we miss something, or add too much, we could make the task of addition too hard or too trivial. This is why we advocate going back to infancy, there the starting state of knowledge should be simpler than at any other age.

The problem of knowing the starting knowledge state does not go away entirely however, determining the innate knowledge of the infant is very difficult. Psychology experiments can be helpful here, in particular the habituation technique. This technique presents an infant with a familiar scene A until the infant is bored of it and looks away, then some other scene B is presented. If the infant continues to be bored by B and looks away, then one can conclude that the infant notices no difference; if on the other hand the infant looks longer at B, then one can conclude that the infant notices a difference. An example of where this can be applied would be showing an infant an object entering and emerging from a tunnel, and then showing the infant one object entering and a different one emerging (younger infants will notice no difference). Furthermore, to support the argument that infants are scrutable (albeit via rather elaborate experiments) an experiment carried out by Bower [1] is most interesting. He separated two groups of infants, and gave one group special training sessions, which involved having them watch scenes of objects pass inside other objects and re-emerge, for example. Simply viewing these displays accelerated the infants' cognitive development and led to improved performance on manual tasks also. The group that received the training reached the next Piagetian stage; the untrained group had not reached this stage in the same time. This experiment is an excellent example of the kind of psychological investigation which could complement a constructivist AI programme. The experiment clearly identifies the precise experiences which have led to the construction of a particular stage of more advanced world knowledge.

Finally, a strong objection to the approach comes from Minsky [9]; he says that the concept of the "baby machine" is reasonable, but that we do not yet have enough knowledge to build it. In particular, he states that "we do not yet have enough ideas about how to represent, organize, and use much of commonsense knowledge, let alone build a machine that could learn all of that automatically on its own". He also cites McCarthy to support this position: "in order for a program to be capable of learning something, it must first be able to represent that knowledge". Our argument above tackles this objection head-on and states that we must build a baby machine precisely because we do not know how to represent much of commonsense knowledge. AI has plenty of examples of programs that can go beyond the knowledge of their designers. Just as Samuel's checkers program played a better game than himself, so we could hope that a constructivist learner could find representations for knowledge which the designer does not know how to code. Obviously it will be challenging to build a learning program when we do not exactly know the target information to be acquired, but we do know the competence which should be displayed, and we know how the competence should change as a result of a certain experience. The designer will have to provide the initial knowledge and method of organisation; this will be refined in an iterative way during the development of a system to follow a particular development path, such as that outlined for the behaviour of the stick. A particular way of organising knowledge will be trialled, and doubtless shortcomings will be found when the system is not able to make the required knowledge acquisitions, and so the organisation will be tweaked, so that it can progress further along the development path, and so on.

A somewhat facile response McCarthy's statement above would be to state that there are learning techniques such as inductive logic programming or genetic programming which can represent pretty much anything, but this is probably not what he meant; they would take too long to find complex representations if starting without substantial background knowledge. In order to be able to learn something in a reasonable time, a program's existing background knowledge and representational framework should be at least pretty close to what is required for the new knowledge. The Piagetian approach proposes an incremental improvement of the representational framework. This is the main argument of the constructivist, which may be at odds with McCarthy; it is that you do not need to have a representational framework in place which is ready represent all required knowledge; frameworks can be tweaked to accommodate the requirements of new knowledge. An argument can be made by juxtaposing a modern scientist with a human adult from a hunter-gatherer tribe. The scientist has representational frameworks which are not present in the hunter, and the hunter could not immediately learn to apply some new algebraic formula. Yet, after an appropriate educational path is followed, the hunter could acquire the appropriate framework and learn the same knowledge. The point is that the hunter has the required constructivist machinery, and that is sufficient. Minsky *et al.* note that "You cannot teach algebra to a cat" [9], and indeed you cannot teach algebra to a baby either, but as a human continues to acquire more sophisticated background knowledge, then at some point it reaches a stage where it has the necessary scaffold on top of which algebraic knowledge can be constructed. By modelling infant acquisitions we would hope to find a computational model of the infant's learning machinery, including how to organise and use new knowledge, i.e. those mechanisms which a cat is lacking.

3.3 Summarising the Prospects

To conclude this section we will summarise the prospects for the constructivist approach to AI, and the promise it holds. By focussing on recreating a particular development path, we are forcing ourselves to come up with a mechanism which can add to its knowledge autonomously. This forces us to tackle the mechanism of intelligence as a central issue. An alternative (and more common) approach to AI is to try to recreate a particular behaviour or competence; the danger here is that specialist knowledge is coded in to solve that task,

and the the resulting system does not shed any light on general intelligence. The constructivist approach holds the promise of acquiring an understanding of concepts (such as size and weight), and having them grounded in sensorimotor behaviours. Such concepts could then serve as the base on which further concepts would be learned through language.

4 PROGRESS

Despite a sizable body of theory from Piaget and others, there has been relatively little work on constructivism in AI to date, and much remains to be tried out. For example, no AI work to date has attempted to model the infant's construction of space which has been sketched in Section 2.1. We will now review some of the main investigations which have been done in constructivist AI.

An ambitious attempt to model Piaget's description of the acquisitions during infancy is the doctoral thesis of Gary Drescher [7] . Drescher built a program to mimic the mechanism of early Piagetian development, and the way in which the concept of an object is learnt. Drescher's program worked in a 7x7 grid world, with a hand, eye and mouth. The program learnt "schemas" which consisted of a context, an action and a result. For example it learnt that if its current context was "HandInFrontOfMouth", and it took the action "HandBackwards", then it would expect to obtain the result "HandTouchingMouth". After exploration it was able to reliably predict the effects of most of its actions from whatever context it was in. Drescher also included a pair of objects in the world, which the eye could see, and the hand could touch and grab, etc. However, one object moved occasionally of its own accord, thus the schema for grabbing it was not entirely reliable. Drescher introduced the idea of the "synthetic item" to cope with this, the synthetic item could be "on" if grabbing had worked recently, and thus could be used to predict if grabbing the object was likely to work. The "synthetic item" is interesting because the program is starting to learn higher order data item which goes beyond what is directly sensed. This is in effect an abstraction of the raw sensor data which allows predictions to be made about objects in the world, and so it arrives at an extensional approximation of the concept of an object (i.e. it is generally "on" when the object is present).

Chaput's doctoral thesis [3] recreated the achievements of Drescher, and went significantly further. Chaput developed a "Constructivist Learning Architecture" which is based on Leslie Cohen's theory of infant cognitive development [4]; this is a neo-Piagetian theory which provides a little more detail than Piaget did about the required information processing mechanism. Cohen has abstracted, from many studies on infants, a set of information processing principles which apply throughout development and across all domains. These principles state that infants learn to process information at increasingly higher levels of abstraction by forming higher level units out of relationships among lower level units. There is a bias to process information using the highest formed units, unless the input becomes too complex, in which case the infant drops back to a lower level and attempts to refine its abstraction so as to be able to handle the complex information at the higher level. Essentially Cohen's principles describe a strategy for making abstractions; from the masses of raw data the infant need only pay attention to the abstractions it has found useful. Chaput's computational model successfully models some aspects of infant development, in particular the perception of causality. He then applied it to Drescher's microworld, and to a robot learning task. In the robot learning task the robot had to "forage" (i.e. see objects, move towards them, and pick them up). The robot had a vision system which had a 60° viewing angle, separated into five sectors of 12° each. When a blob appeared in one of these sectors, then a binary sensor item was set to true. Chaput used Cohen's information processing principles to construct synthetic items in an efficient way. In the foraging task, Chaput's system came up with many interesting and very useful synthetic items, which led to very efficient performance on the task. To compare Chaput's system with the system that would be required to learn the behaviour of the stick above, many of the required elements are there, but more machinery would be required to be able to make analogies between schemas. Chaput's robot sensors are too simple to allow such analogical machinery to be useful (the robot could not see one object in front of another for example).

Cohen et al. [6] provide a quite different approach to coding Piagetian schemas, which is somewhat more complicated, with action schemas containing "maps". These maps can represent a space with dimensions of distance and velocity, for example, and they can record the activation of a schema as a trajectory in this space. Their system can learn "gists" which are compositions of action schemas for certain tasks. This has been successfully applied to learn behaviours in a simulated world, for example a creature learns to sneak up on, and catch, a cat. The schemas learnt have also been transferred to similar situations in slightly different scenarios. This mechanism might well be applied to the developmental path we have outlined, in this case we would like to add to the mechanism so that it could see analogies between similar schemas or similar gists. This would allow a partial match to be found between two gists, and a conjecture to be made that unmatched aspects might also be similar.

A final related work which is worth citing is by Buisson [2]. This work learns to recognise the rhythm of a piece of music. The system features an active type of perception which attempts to match the rhythm being played by playing its own schemas to synchronise with it. It uses an evolutionary algorithm which starts with its own basic schema and generates mutated schemas, some of which will match with the target rhythm, and some not. Those which match will replicate and those which do not will die. Buisson's work seems to be the only computational investigation of Piaget's theory which really takes Piaget's idea of "assimilation" seriously. By assimilation Piaget means the way that experiences in the environment can be matched to known schemas. Buisson takes this seriously by acknowledging that the rhythm which is being played cannot be simply copied if it is not known already; the program must be active in conjecturing a variation on rhythms it knows to see if this might match the rhythm being played.

5 CHALLENGES

We will now outline some of the major challenges that need to be overcome to successfully exploit constructivist theories in AI.

(1) We need to find plausible development paths to copy. Piaget gives us some sketchy accounts, and much work will be required to flesh out the details of these. Research on infants could be particularly helpful here, in particular the type of study conducted by Bower [1], as discussed in Section 3 above. For example, we could investigate what kind of training would accelerate an infant's development of the behaviour of the stick, this would allow us to identify the experiences which infants profit from to develop the behaviour. We could also investigate what types of landmarks are being used by the infant, by investigating if their absence affects the behaviour. Such studies are very resource intensive unfortunately, but the idea that we could really settle some of these questions with infants is exciting.

(2) Given a particular development path, we need to find a computational model to explain it. Particularly tricky issues here include developing appropriate analogy finding mechanisms, finding appropriate representations for schemas of knowledge, and organisation among schemas (in particular sub- and super-schemas).

This challenge may not even be achievable in an AI system. Finding a development path in infancy would show that it is possible in the case of the infant, but whether this extends to an AI system (simulated or embodied robot) with different sensors and effectors is not clear, and will need to be investigated.

Assuming that it is possible, we can foresee the following type of iterative progress: the system will successfully reach a certain stage in the target development path, and will be incapable of making the next step. It will then be necessary to add innate knowledge or abilities to the initial infant, to see if that can allow it to go further. This is to be expected to some extent, because there may be some aspects of the infant's innate mechanism which have been present since the beginning, but just had no chance to express themselves before a certain level of knowledge was reached. Thus our system could reach a simpler level of knowledge without these abilities and then find that something is missing. The danger however is that, in helping the artificial infant to get to a particular milestone, we may put in innate abilities which could in fact have developed by themselves.

There are a number of examples in Piaget's theory of "new" behaviours that are expressed at a certain age, which seem to appear from nowhere; however, a detailed reading of his theory typically gives an account of how they are developed by the same mechanisms that have been at work since birth. An example of this is experimentation; the infant seems to suddenly start experimenting with the parameters of actions, and varying them, at about twelve months. However Piaget accounts for this by explaining that it arises because he now has so many schemas to recognise effects of actions; if a slightly different result is accidentally produced by an action, this difference will be salient to the infant, and the infant will try to ferret out the parameters of the action which can cause it. Thus the development of schemas which has been happening all along can account for the "sudden" emergence of experimentation.

It is also important that the world in which the artificial baby develops is sufficiently rich to allow the baby to develop all the behaviours we require. Otherwise it might fail to achieve some milestone not because of any deficiency in the innate mechanism. Getting this right requires a deep understanding of the development path we are trying to follow, and the nature of the knowledge which should be acquired; it will also doubtless require some trial and error.

(2.1) The challenge is really to explain the development with a minimal mechanism. We could say that the challenge of the programme as a whole is to show how qualitatively different behaviours can emerge from the continuous operation of a single mechanism (we are proposing that this mechanism be researched by trial and error with AI systems). We need to be careful about adding a new ability to the innate mechanism, when it might be possible to make the new ability emerge by providing a sufficiently rich world, or an appropriate developmental path which allows the existing mechanism to develop it.

(4) As the number of schemas grows, combinatorial problems will arise. It will not be possible to search for correlations between all schemas in order to find relationships to explain new phenomena; schemas will need to be searched selectively, and to be organised in some fashion. This is one of the problems that Minsky cited in his criticism of the baby approach. Possible solutions might be found in Cohen's [4] principles.

6 CONCLUSION

We have claimed that there is much value to be gained by applying constructivist theories to Artificial Intelligence. We advocated a research methodology which would set the modelling of a developmental path as a goal, rather than the achievement of some particular competence. This would force the research to investigate the constructivist mechanism itself (and hence analogy, among other things). It is hoped that this would lead to systems which are more adept at general tasks, when compared with classical AI systems (which achieve competence on constrained tasks).

We outlined a particular development path which might be attempted by the constructivist AI approach. This path followed the infant's laborious construction of space by analogy, and grounded in known actions. Because this construction is a laborious process, it must bring some benefit to be considered worthwhile. To justify this we explained that a construction which is built in this laborious way is very useful for a system that needs to act in the world, because its perception of the world is now built from more primitive actions which it knows. With this argument we hoped to point out the value of Piaget's theory, by showing why AI might do well to model this somewhat idiosyncratic development path.

We also saw that the main element currently missing from related work is the ability to find analogies among existing knowledge, and so to conjecture more elaborate models of the world. In this respect there is much more to be exploited in Piaget's theories.

Acknowledgements: Special thanks to Nir Oren for comments.

REFERENCES

[1] T. G. R. Bower, *Development in Infancy*, San Francisco : W.H. Freeman, 1982.

[2] Jean-Christophe Buisson, 'A rhythm recognition computer program to advocate interactivist perception ', *Cognitive Science*, **28**(1), 75–87, (February 2004).

[3] Harold Henry Chaput, *The constructivist learning architecture: a model of cognitive development for robust autonomous robots*, Ph.D. dissertation, Artificial Intelligence Laboratory, The University of Texas at Austin, 2004. Supervisors: Benjamin J. Kuipers and Risto Miikkulainen.

[4] L. B. Cohen, 'An information-processing approach to infant perception and cognition', in *The Development of Sensory, Motor, and Cognitive Capacities in Early Infancy*, eds., F. Simion and G. Butterworth, 277–300, East Sussex: Psychology Press, (1998).

[5] L. B. Cohen and C. H. Cashon, 'Infant perception and cognition', in *Comprehensive handbook of psychology. Volume 6, Developmental Psychology. II. Infancy*, eds., R. Lerner, A. Easterbrooks, and J. Mistry, 65–89, New York: Wiley and Sons, (2003).

[6] Paul R. Cohen, Yu-Han Chang, Clayton T. Morrison, and Carole R. Beal, 'Learning and transferring action schemas', in *IJCAI*, ed., Manuela M. Veloso, pp. 720–725, (2007).

[7] G. L. Drescher, *Made-Up Minds, A Constructivist Approach to Artificial Intelligence*, MIT Press, 1991.

[8] M. M. Haith, 'Who put the cog in infant cognition: Is rich interpretation too costly?', *Infant Behavior and Development*, **21**, 167–179, (1998).

[9] Marvin Minsky, Push Singh, and Aaron Sloman, 'The St. Thomas Common Sense Symposium: Designing Architectures for Human-Level Intelligence', *AI Magazine*, **25**(2), 113–124, (2004).

[10] J. Piaget, *The Origin of Intelligence in the Child.*, London: Routledge & Kegan Paul, 1936.

[11] J. Piaget, *Play, dreams and imitation in childhood*, London: Heinemann, 1945.

[12] J. Piaget, *Construction of reality in the child*, London: Routledge & Kegan Paul, 1954.

[13] M. Ridley, *Evolution*, Oxford: Blackwell, 2003.

[14] A. Sloman, J. Wyatt, N. Hawes, J. Chappell, and G. J. M. Kruijff, 'Long Term Requirements for Cognitive Robotics', *Proc. Cognitive Robotics '06 Workshop, AAAI'06*, (2006).

Social Robotics and the person problem

Stephen J. Cowley[1]

Abstract. Like computers before them, social robots can be used as a fundamental research tool. Indeed, they can help us to turn our attention from putative inner modules to thinking about the flow and emergence of human intellectual powers. In so doing, much can be gained from seeking solutions to MacDorman's *person problem:* how can human bodies – and perhaps robot bodies – attune to cultural norms and, by so doing, construct themselves into persons? This paper explores the hypothesis that social robots can be used to ask fundamental questions about the nature of human agency.

For social robots to live up to their name, the focus needs to fall on functional co-ordination and co-action. This enables one to link research on how today's robots function as social mediators with engineering approaches that explore both how understanding can be hard-wired, how this influences the cultural ecology and, perhaps, in designing robots that can discover how we enact values. To do this new kinds of collaboration need to be established. The key theoretical question is whether, in becoming persons, humans depend on embodiment alone or, as suggested here, intrinsic motive formation enables them to discover the distributed forms of embodiment favoured by culture.

1 INTRODUCTION

Since social robotics is in its infancy, no-one knows what impact these machines will have. Especially in the West, their potential is apparent to few. I felt it – and I mean felt it – when I went to Osaka to talk about infant development. There I was introduced to androids; my hosts suggested that, in realizing their power, I might help by proposing an engineer friendly model of human behaviour. What follows is, in part, a polite refusal. Rather than build robots that simulate what psychologists *say* humans do, social robots can free us of mentalist fantasies. To do this, I argue, they must be recognised as a research tool. Their value arises in investigating the real-time flow and changing results of joint action. Eschewing appeal to psychological competencies, engineering design can thus co-emerge with observational studies of humans and systematic investigation of how a cultural ecology adjusts to social robots.

Co-action can be defined as occurring when "one agent's action is influenced by or occurs in the context of another agent's– and together they do something that is not fully attributable to either one alone" [50]. As explained below, robots with this capacity can *simulate* understanding of what happens between people. Given that they already produce social affordances, this opens many new applications. This paper however, focuses on other questions. Just as computers showed us about cognition*,* it is argued, robots can be used to understand human forms of minded behaviour. As MacDorman and Ishiguro emphasise, robots can be a test-bed for fundamental research [33]. Whereas computers clarified thinking about *mind,* robots enable us to formulate and test hypotheses about human agency. By contributing to social encounters, they can throw light on how physical and cultural resources impact on what we think, feel and do. By asking bold questions, robotics may lead to syntheses that link the work of engineers, psychologists, philosophers and social scientists.[2]

2 BEYOND COMPETENCIES

Since Chomsky's work in the 1960s, the engineer's universals have dominated models of human powers. Notoriously, even language is often associated with an innate module. As a result of a *separatist* approach, cognitive science has faced difficulties – the frame problem, symbol grounding and, of course, the 'hard' problem of consciousness. AI need not work thus. In the early days, Turing and Craik sought to focus the mind sciences – not on competencies and tasks – but on 'intellectual capacities.' Although we may understand these no better than 50 years ago, we now know that they exploit actions and artefacts as well as the brain's equivalent of *software*. In 1950, Turing's imitation game presupposed machines that computed (seemingly) intelligent responses to typed word-strings [46]. Such devices, he thought, could imitate human word-use. This proved to be illusory. Computers calculate better than us, mimic chess-playing and provide practical applications for, say, rudimentary vision. However, they cannot understand word-strings. Without bold thinking, therefore, meaning would never have been traced to self-organizing capacities which integrate neural, bodily and material resources. We would not have realized that, somehow, embodiment grounds semantics.

In the *Nature of Explanation,* Craik focused on our capacity to come up with *objectively valid* models [16]. Meaning depends on *external processes* that prompt us to reason with public symbols. In bridge-building, for example, we agree in our judgements about what symbols mean. Of course, while Craik posited that semantics were exclusively managed by the brain, today we know that this plastic system uses external resources to self-organize [22]. Pure software models are thus giving way to models that leave space for cognitive *deeds.* As this acronym suggests, intelligent activity is dynamic, embodied, embedded, distributed and situated [49]. In popular metaphors, natural born cyborgs use material symbols in a cognitive niche [7, 8]. More sober rhetoric invokes embodied embedded cognition (e.g. [52]) or, perhaps, enactivism (e.g. [43]). Finally, for humans, Hutchins' work is seminal. To make valid judgements about a

[1] Psychology, University of Hertfordshire, UK; University of KwaZulu-Natal, Durban, South Africa. Email: s.j.cowley@herts.ac.uk

[2] We explore the *origin of selves* – not at the primordial level of self-other boundaries – but in terms of how we become spinners of narratives [20, 21].

ship's position, naval personnel integrate cultural and physical resources with co-action [29]. Cognition is *culturally distributed*.

Once we look beyond competencies, we discover a domain of culturally distributed meaning. How, then, do symbols guide us to judgements? Instead of positing a language module, we can regard our linguistic skills as the product of self-organizing brains and bodies [22]. Gradually, as we integrate action, perception and external resources (including language), brains set up bio-cultural control systems or *selves*. Ross and Dumouchel say:

> Biological systems of the *H. sapiens* variety turn themselves into people—socially embedded teleological selves with narrated biographies in terms of these very beliefs and desires—by taking the intentional stance toward themselves. They can do this thanks to the existence, out in the environment, of public languages that anchor their interpretations to relatively consistent and socially enforced rules of continuity.... [T]hey are incentivized to narrate themselves as coherent and relatively predictable characters, and to care deeply about the dramatic trajectories of these characters they become...[People] are partly constituted out of their social environments, both in the networks of expectations that give identity to them as people, and in the fact that the meanings of their own thoughts are substantially controlled by semantic systems that are collective rather than individual. They are thus not identical to their nervous systems, which are indeed constituted internally. [40] (pp. 264-265)

MacDorman pursues this view by posing the *person problem* [34]. Eschewing the metaphor of embodiment, he asks, "How could human bodies – and perhaps robot bodies –attune to norms and, by so doing, construct themselves into persons?"

3 THE PERSON PROBLEM

Do human bodies use cultural ecology to become persons? To test this hypothesis our focus falls on – not modules – but *cognitive integration*.[3] Leaving behind the view that something like software could ever be sufficient to explain intellectual powers, we take a *one system view* [30]. Instead of relying on separating out aspects of nature, we ask how what we do, feel and say arises in bodies-in-the-world. Indeed, as Blackburn suggests, there is no mystery to what is distinct about human agency [5]. To know what someone is doing, we need to see their movements as *expressions of intention and purpose*. However, we lack a viable model of how this is done. Davies writes:

> What kind of agent are we? My answer is twofold. First we do not know. At this point in the history of inquiry, traditional notions of agency are dead or dying and their replacements have yet to be born or yet to reach maturity. Second, although we are in a period of conceptual transition and thus we have no developed concept of human agency, we do know what kind of agent we are not." [20] (p. 39)

Once we abandon mentalism, our powers can only be traced to selection mechanisms. Further, given the co-evolution of nature and culture, human forms of agency must be grounded in action and perception. As Davies argues, biological agents use norms – nonsymbolic coding – at all functional levels. We must rely on these to become agents who describe our doings in terms of symbolically represented plans and goals. Their basis must arise from "causal or structural capacities which contribute to the exercise of some larger systemic capacity in the larger systemic view [19], p. 4. Nonsymbolic coding, it is argued, operates from the level of organism right down to the cell.[4]

Norm-based biological models provide a basis for exploring how we orient to norms, use them to constrain our dynamics and, by so doing, higher levels of control. They enable us to think about the nature of agency. For this reason the person problem falls in line with the tradition of Craik and Turing. For, in asking how we become persons, we pose a constructive question. Instead of saying why this is difficult, we ask how biology prompts us to use external resources as a basis for developing ways of controlling minded behaviour. Self-organizing processes link cultural resources with the body such that we develop intellectual powers. As a result, we occasionally come up with *objectively valid* judgements. Can robots mimic this? Can they use a changing context to move in ways that humans find appropriate*?* On this view, flexibility is more important than definition. Indeed, as Wittgenstein saw, even language depends on 'agreement in judgements.' If biological norms allow us to classify particulars (using what Anderson calls *action-guidance representations* [1]), these track aspects of the physical world. Humans, however, also draw on the practices of fellow agents. Given norm-based experience we redeploy our representations as perception becomes associated with how other people are likely to react. Humans thus unite propensities for statistical learning with norm-oriented motivations. Objects, situations and events are constituted, at least in part, by attitudes. Even infants set off contingent responses whose value is monitored and, when rewarded, associated with cues used in anticipatory action [41, 44, 12, 10]. Can robots use human co-ordination in coming to self-construct what (naïve) humans regard as judgements?

4 INTERACTION, FUNCTIONAL-COORDINATION AND CO-ACTION

Although dominant in human life, co-action is not easily characterised. Indeed, in folk psychology, it is often called *interaction*. While readily separated from acting on nonliving things (whose properties produce a nexus of relatively predictable relations), it is more difficult to tell apart from behaviour between living organisms. Recently, interaction has been defined with respect to co-ordination [23]. In encounters between organisms, this comes to constitute an *autonomous* interactional domain. Accordingly, routine co-ordination – when starlings flock or people co-ordinate finger movements – contrasts with *functional co-ordination.* In such cases coupling demands new kinds of integration. Characteristically, this results from disruption. For example, a stickleback may *fail* to perform a mating move or, when walking down a corridor, we may find

[3] Menary opposes *cognitive integration* to the view that cognitive abilities are solely, or essentially, neural. In minded behaviour, neural and bodily processes are integrated with external vehicles [37,].

[4] These are 'nonsymbolic' in that they are not reducible to physical tokens that are manipulated in accordance with a syntax; however, in Pattee's sense, they are symbolic [39]. They are rate-independent parameters with a selection history that is inseparable from dynamics. Interestingly, this meshes very closely with Barbieri's view of biosemiosis [4].

ourselves about to collide. In such case, both parties may adjust or co-ordinate functionally. Events beyond the body influence what happens and, in the human case (at least), how each party feels. The achievement of De Jaegher and Di Paolo is to show that this emergent organization can be defined with respect to both mutual influence between individual actions and the concurrent influence of relational dynamics [23].

For Wegner and Sparrow *co-action* occurs when "one agent's action is influenced by or occurs in the context of another agent's – and together they do something that is not fully attributable to either one alone" [50]. If we compare this with functional co-ordination, we can be more precise. Whereas De Jaegher and Di Paolo write with biological norms in mind, in human life events are also affected by the cultural nexus [23]. This level of organization enables us to exploit co-action.[5] Human agents use a history of coping with contingency together with how experience prompts each party to expect the other to react to what is possible. As a result, we develop anticipatory skills by evaluating the 'something' that results from co-action. In contrast to association-based learning, this uses – not physical cues – but situation-based evaluations.[6] Since these set off learning, co-action can channel development. Returning to Wegner and Sparrow, the *something that occurs* underpins agreement in judgements. This unites the child's intuitive grasp of the value that adults attribute to co-ordinated movements. One party (at least) construes the co-ordinated movements in terms of norms (that can be reported as reasons). Cultural know-how induces them to react to these as expressions of intention and purpose. This attribution has a remarkable outcome. It ensures that, for all parties, a well-timed movement can realize *local values*. Indeed, as Gibson first suggested, this may be the basis of social learning [26, 28].

In illustration, one can consider an event described by Cowley and colleagues [12]. In this, a 14 week old baby interacts with her Zulu speaking caregiver. Strikingly, the baby seems to know that her mother *really* wants silence. When she does so, the reward is a tease and a beaming smile. Co-action arises as the mother captures the baby's attention with salient, rhythmical in-your-face hand-signals,while using harsh vocalizations to get the baby to behave (these are musical and verbal). Since the event depends on what happens beyond the body, the baby acts neither automatically, imitatively, nor by drawing on any kind of universal. Rather, using experience, both parties orient to the cultural context. To co-act by realising a hoped-for value, the baby uses a pattern that, in Zulu homes, often serves in urging a baby to *thula* (fall silent). Indeed, given local beliefs, caregivers often obtain respect without offering physical rewards (e.g. hugs or physical contact). Where the infant does as hoped for – shows *respect* – this is seen as something special. Acting together with the caregiver, 'understanding' is enacted as *both* bodies orient to local norms. The event makes sense because the baby enacts a cultural value (respect). This is more complex than functional co-ordination because the adult habitually realises the value (with variations): she enacts – not movement-types – but a culturally encoded intention (to get her baby to *thula*). Given experience, the baby discovers values realising behaviour. As in this case, this depends on predispositions to use adult response to body-based contingencies. Trevarthen has long argued along these lines [44]. Join affect enables co-action (or *intersubjective behaviour*) to exploit – not just bodily attunement – but neural development based in *intrinsic motive formation* [45, 11].

5 THE TRUTH ABOUT SOCIAL ROBOTS

On the face of things, we might expect social robots to engage in culturally-based co-action or, at least, functional co-ordination. While this may occasionally occur, research has tended to focus on individual competencies. Even social activity is often modelled – not as co-action – but as *imitation* or, at least, as reducible to *syntactic processing*. In such cases, one set of behaviours – or formal patterns – are mapped onto another. In contrast to both functional co-ordination and co-action, such systems fail to go beyond given information or, indeed, cope with even simple disruptions. This logic is defended by Dautenhahn: "Life and intelligence only develop inside a body [which is] adapted to an environment in which an agent lives" [18]. Where the organism is taken to consist in what lies *within* the skin, functional co-ordination seems mysterious. Indeed, models that separate the individual from the world are bound to underplay real-time activity. Autonomy has to be conceptualised in terms of individuals who are "embedded, coupled and linked to a social context" [18]. Placing the living on the inside, context is idealised around external invariants that, it is hoped, can be discovered by *complete systems* [18]. Up to the present, then, social robotics has adopted what Järvilehto decries as the two-systems view [30]. For example, in Breazeal's work, what matters is mainly whether the robot supports the model that the engineer has assigned (if it displays a competency) and secondarily whether this can be successfully displayed in the 'assigned interaction scenario' [6]. By contrast, if the kinds of co-ordination displayed by the living are the basis for social life, we are in for a surprise.

Emotion-showing robots such as Kismet are not social. They manifestly fail to display the adaptive, flexible behaviour that is typical of living systems (even at the level of a single cell). This is because they have been designed – not with an eye to biology – but to replicate competencies. Building on folk views of the individual, they are expected to contribute to the doings of heterogeneous groups of humans and/or other robots. The putative competencies are expected to scale up in ways that will eventually give us 'embodied agents' who perceive and interpret the world in ways that enable them to 'recognise each other and engage in social interaction' [18]. Ideally social robots will becomes individualised agents who, 'on the basis of their own experience' can 'explicitly communicate and learn from each other' [18]. Since they – and, by implication, we – only simulate sociality, the approach introduces many puzzles. For one thing, interaction seems to be an inadequate basis for explaining either sentience or social perception. With current

[5] In contrast to the approach taken here, it is possible to take the view that there is a continuum from simpler to complex types of functional co-ordination [23]. While this eliminates appeal to intrinsic motive formation (see below), I find it hard to see how it can give an agent the power to realise culturally-based values. This seems to demand the more extended kind of embodiment defended here.

[6] Ariely & Norton,critique the view that actions are driven by the sum of hedonic aspects [2]. Actions-in-the-world impact on future behavioural trends or 'contexts are endogenous to decisions' In this kind of *'self-herding'* past behaviour enables us to derive an arbitrary set point -not necessarily based on hedonic input. If prompted, this functions as input to action and inferences-based-on-action. It thus influences (but does not cause) calculations (and self-reports)

technologies, machines lack the sentience that shapes individual experience and, without this, it is hard to see how they could acquire the 'symbols' needed in explicit communication and learning. Indeed, one runs into all the problems that arise from reducing human powers to the competence models that are the mark of mentalist tradition.

The person problem presents social robots as part of a single system or cultural ecology that includes humans. Instead of highlighting individuals, the question becomes how they can be integrated into our cultural world by using co-ordination and, ideally, developing a manifest sensitivity to social norms. To do this, of course, social robots will have to mimic bio-cultural aspects of agency. Using cognitive integration, like human infants, they need to detect cultural norms that can be used to reshape their control systems. Given what we know of infants, this depends on physical objects, interactional relations the evaluation of (joint actions) and how contingencies map onto local norms. While lacking space to discuss these matters, today's synthetic models already use such resources. For current purposes, however, I highlight another issue raised by the person problem. To understand our ecology, we can study robots *in the wild*. As Hutchins did in studying naval practice, we can ask how agents use social robots as resources that, among other things, lead them to agreement in judgements [38]. Thus, rather examine how robots influence us. Just as was done for computers, we can take on board the idea that a primary use of social robots is, at this stage, as tools that can be used to generate and test hypotheses about minded behaviour.

While not designed to use co-ordination and social norms, we *feel* that social robots are social. Indeed, they are designed to make us treat them more like living dolls than fancy computers. As is widely attested, even simple robots become *relational objects* [47] that set off anthropomorphic reactions. Larger, more robust machines can have a dramatic impact on a cultural ecology. This appears dramatically when Robovie, a Japanese-designed social robot, is introduced to classroom environments. Even if the machine's interactional routines are governed by software, humans attempt to set off co-action [42, 38]. In longitudinal trials, the classroom ecology alters in dramatic ways. Further, while initial interest may fall off, the robot evokes high quality interactions and, for some of the children at least, these are maintained over time. The outcomes have been extensively studied: work includes several case studies, the micro-coding of response to robot behaviours, attention to global variables labelled sociability and familiarity and, recently in *content behavioural analysis* [15]. In this latter approach, the focus falls directly on content-patterns that arise as children structure their co-action around the machine. Thus, as in traditional content analysis, human-human-robot analysis is described in terms of recurring themes.

6 LEARNING FROM HUMAN CO-ACTION

Robovie functions as a social mediator. While long recognised that robots *mediate contact between children* [51], the idea of a social mediator has a long historical tradition [38]. Using micro-analysis, Nabe and his colleagues show that robot behaviours can take on language-like meaning or, in Vygotsky's (1981) terms, became *mediational means* [48].[7] They function as psychological tools by allowing human agents to access a nexus of cultural norms. Indeed, as Hegel and Marx argued, such means locate the subject in a historical context. In the classroom, mediational means function to spur study and play that enable children to develop both individual skills and co-action routines. While many robot behaviours elicit, at best, stock responses, others become increasingly valued [38]. In short, while most seem pointless (and are ignored), others set off enjoyable events. This is not surprising. Babies too ignore most of what happens but, given certain sorts of attention, make co-action attempts. Unlike baboons (but perhaps not chimpanzees) much of our behaviour is directed –not at goals –but through experience of external resources. Humans act epistemically [14]. We act in world-directed ways by seeking out interesting effects. In the classroom, this happens because children anthropomorphise the robot by seeking to set off social action. Indeed, were the robot human, what they do would induce rewarding response based on shared orientation to local norms. The robot, however, does not perceive fine human movements – let alone their social meaning. For this reason, it is incapable of either functional co-ordination or, by extension, evaluation of co-action attempts. Given the classroom ecology, however, children respond to *each other*. Thus certain robot-directed co-action attempts become valued [29]. Ways of orienting to robot behaviours become contexts that can be used to set up achieving social effects – one's based on co-action. Using a semantic nexus associated with a movement, children learn from robots.

Can robots learn from humans? Of course. However, machine learning tends to be conceptualised in terms of specific tasks. Typically, what is modelled presupposes a competence that is the theoretical descendent of concepts like the *mental lexicon.* Then, using this model, the system learns tasks such as associating sound-patterns (words/phones) with images or acoustic patterns. To address the person problem, however, human agency can be traced to how, in real-time, we orient to social norms. Instead of evoking competencies or minds, we can focus on how acting jointly, we come to realise values. Are words be learned this way? Elsewhere Cowley presents this as the basis for *human symbol grounding* [10]. This, then, is a separate issue from learning to use physical invariances to track salient aspects of the world. In other words, as developmentalists often argue (e.g. [32]) rather than relying on linking mind to world, we rely on linking what is perceived to expectations based on the experience of persons. This leads to simplification. Above all, objects can be recognised with respect – not to invariances – but culturally salient aspects. Thus babies come to perceive, say, moods, possible actions and the likely arrival of dinner. To simulate functional co-ordination, therefore, it would be possible to rely exclusively on programming. Robovie, for example, might use sensor input to calculate the laughter-frequency of individuals and, using rules, produce canned laughing (or other behaviour) to join in with them. There is little doubt that this kind of integrational process would have a massive impact on the classroom ecology.[8] While there is

[7] Vygotsky listed counting, mnemonic techniques, algebraic symbol systems, works of art, schemes, diagrams, maps, mechanical drawings, all sorts of conventional signs and so on [48] p. 137.

[8] Robovie uses pseudo learning. After a certain number of interactions with the device a child is rewarded with a secret. This is an effective technicque that has striking classroom effects [13].

nothing intelligent about this, much the same goes for human laughter. This too has an automatic basis. What is striking, then, is how we are able to draw on a cultural nexus in learning to control when – and how – to laugh. That, however, is complex: it is already a clear case of co-action.

Programming could also be used to simulate co-action. One could hard-wire biases that prompted the robot to identify a subset of behaviours. Second, with more hard-wiring, these could be mapped onto cue-defined context-types. In the laughter example, the robot could distinguish this from other vocalizations and relate both to the t-behaviour (teasing) that recur in the Japanese classroom [15].[9] Using cues like whether, at time *t*, more than one child laughs – or all laugh – follow-up response could be varied. Third, response-by-laughing could, as it is in the children, exploit capacities for synchrony, simultaneity and setting up sequences. This would change their encounters. However, less dramatic examples might have more powerful effects. Given the (rather stupid) ways in which children rub and tap different sensitive parts of the robot, interesting simulations of co-action could be induced by timing vocal patterns that simulated affect which were systematically associated with kinds of touching. There is little doubt that children would be enthusiastic. Contingencies and co-ordination are at the heart of human social life. In proposing this approach to social robotics, therefore, the focus falls on using robots as test-beds to how, given these physical resources in this culture, humans construct co-action.

7 ADDRESSING THE PERSON PROBLEM

Could design replace simulation? Might machines use movement to detect affect and simulate the motive formation which underpins co-action? If so, we might make progress with the person problem. The difficulty, if there is one, lies in identifying – not physical patterns – but variable contingent cues associated with abstract norms. A machine would have to *discover* something as does the Zulu baby. If control-side, the agent is in state X, it must calculate that world-side cues *mean* either (a) inhibit; or (b) do not inhibit. Depending on a decision, a baby is likely to go into either state Y (e.g. hurting less) or state Z (e.g. crying more). Finally, this will influence longer-term rewards (maybe being abandoned; maybe being picked up). Of course, such decision making does not develop from scratch: like the baby, the robot would use experience to form motives while using biased sensors to predict likely up-coming rewards. This may be difficult. Nonetheless, a machine that did this would meet several of the proposed benchmarks: it would function as an embodied agent that simulated *experience* to *perceive and interpret* a small aspect of *the world*. While this would not be sufficient for social perception or the development of capacities for fully-fledged interaction, such an agent would learn from how humans co-act. Crucially, it would act – and not act – by integrating its own states with human propensities to reward certain norm-based co-actions. It would function as part of the cultural ecology – as a one-system model or an an integrated mediational means.

Much empirical work points towards ways of approaching the person problem. What is more challenging is that the approach demands ways of integrating research based in, for example, simulations, ALife and robotics. New kinds of collaboration are required. By way of illustration, let us sketch recent progress in ALife and Android Science. Thus even artificial environments can be used to detect variable contingent cues that are needed for functional co-ordination [3]. While independent of historically-derived norms, the world serves in shaping an agent's powers. Given central nervous systems (or computation), variable contingent cues can take on many functions.[10] To develop into simulations of co-action, different sensorimotor means would be needed in integrating activity. Here android science provides machines that not only elicit culturally appropriate response, but rely on a human subject's (norm-based) expectations. For example, gaze-behaviour is human-like if – and only if – people believe that the android is under human control [35]. Gaze is a co-actional resource: looks function in the context of another agent's gazing – and, given a nexus of norms, set off something which can be attributed to neither party. In human life, this strange property is central to both mother-infant interaction and, say, flirting. Going beyond the information given is equally compatible with the emerging view that language – far from depending on symbolic competencies – derives from the physics of expressive co-action. Androids, of course, are the ideal test-bed for exploring how expression is integrated across modalities and between parties [33]. Even today, they could be programmed to prompt humans to use gaze in jumping to conclusions that are, in many cases, unwarranted[11]. It is an empirical question whether they could discover how gaze contributes to co-action. Strikingly, roboticists may be in a strong position to give new insight into how gaze is integrated with both other expressive dynamics and verbal patterns.

8 TOWARDS CO-ACTING ROBOTS

It is so hard to overthrow mentalism that even in the embodied, embedded tradition some invoke extended *mind* [9]. As Gibson suggests, this may be because human learning relies on meaning and value [26, 28]. Indeed, in an early paper, he argues that a viable learning theory will draw on – not behaviourism – but social psychology. Only then can learning be seen, not as the basis for social life but, rather its consequence. As emphasised above, humans act to realize values. In focusing on co-action, we propose that human bodies construct themselves into persons by using the movements of others. The trick, however, lies in being designed to *take these* as expressions of intention and purpose. For this to be possible, human babies exploit *intrinsic motive formation* [45]. As a result they can use not only disruptions that lead to functional co-ordination but also cues that mark historically derived cultural norms. Gradually, they become biocultural agents. We construct ourselves into persons by learning how to use the doings of other in realizing values.

It has not been demonstrated that co-action cannot derive from functional co-ordination. *That* is an empirical question. However, if functional co-ordination is sufficient for values realizing behaviour, even intellect can be traced to embodiment.

[9] In the current model of content-behavioural analysis, three of the fifteen themes that can be associated with teasing.

[10] This could be construed as a way of talking about affordances.

[11] While extreme gaze sensitivity is especially prominent in psychotics, it is rare only among people in the autistic spectrum [17].

On the view presented here, however, co-action introduces a new level of complexity. Just as functional co-ordination depends on mutual influence between individual actions and the concurrent influence of relational dynamics, co-action exploits norms that are realised by material acts that carry historical patterns. In this way real-time relations can be linked with the slow dynamics associated with habits. Values can be realised by orienting to a cultural nexus of norms or the *interaction order* [27]. Slowly we develop shared models of interaction-types. By the age of 4 or 5 children get a sense of what, for example, distinguishes a doctor from a patient. Cultural norms associated with verbal patterns prompt thoughts about which people agree. Given their growing sense of what is expected, children come to act strategically. With agreement in judgements, their views come to have some objective validity – "Doctors make you stick your tongue out." Children become role-playing agents whose preferences draw, in part, on experience and, in part, on the rewards of self-display. In learning about strategic signalling, we discover how to play and, gradually, turn into persons. We become skilled with a range of practices and, in the end, may learn the tricks and practices that, as in Craik's [16] example, enable us to interpret symbols as the plans for a bridge. This depends, not on inner competencies, but discovering the values needed by a bridge-building engineer. By posing the person problem, we suggest that much can be gained by tracing social skills to co-action: this, it is hypothesised, is nature's trick for getting us to treat visible and vocal gestures (speech) as expressions of intention and purpose.

REFERENCES

[1] Anderson, M. & Rosenberg, G. Content and Action: The Guidance Theory of Representation. In: *Evolutionary Biology and the Central Problems of Cognitive Science*, D. Smith (Ed.) *Journal of Mind and Behavior,* 2007.

[2] Ariely, D. & Norton, M. How Actions Create -not just Reveal Preferences, *Trends in Cognitive Science*, 12/1: 13-17. (2007)

[3] Auvray, M., Lenay, C. & Stewart, J.. The Attribution of Intentionality in a Simulated Environment: the Case of Mininalist Devices. In: *Tenth Meeting of the Association for the Study of Consciousnes*s. Oxford, UK, 23-26 June, (2006).

[4] Barbieri, M. Is the Cell a Semitioc System? In: *Introduction to Biosemiosis: the New Biological Synthesis.* Dordrecht: Springer, M. Barbieri (Ed.) pp. 179-208. (2007).

[5] Blackburn, S. *Ruling Passions,* Oxford: Clarendon Press. (1999).

[6] Breazeal, C. Towards Sociable Robots. *Robotics and Autonomous Systems*, 42/3-4: 161-175. (2003).

[7] Clark, A. *Natural Born Cyborgs: Minds, Technologies and the Future of Human Intelligence*. Oxford: Oxford University Press. (2003).

[8] Clark, A. Language, Embodiment and the Cognitive Niche. *Trends in Cognitive Sciences*, 10/8: 370-374. (2006).

[9] Clark, A & Chalmers, D. The Extended Mind. *Analysis* 58(1): 7-19. (1998).

[10] Cowley, S.J. How Human Infants Deal with Symbol Grounding. *Interaction Studies, 8.1: 83-104.*. (2007).

[11] Cowley, S. J. The Codes of Language: Turtles All the Way Up? In *The Codes of Life*, M. Barbieri (Ed.) 319-345. Springer, Berlin. 2007.

[12] Cowley, S.J., Moodley, S. & Fiori-Cowley, A. Grounding Signs of Culture: Primary Intersubjectivity in Social Semiosis. *Mind, Culture and Activity, 11/2:* 109-132. (2004).

[13] Cowley, S.J & Kanda, T. Friendly Machines: Interaction-Oriented Robots Today and Tomorrow. *Alternation,* 12.1a: 79-106. (2005).

[14] Cowley, S.J. & MacDorman, K.F. What Baboons, Babies and Tetris players tell us about Interaction: a Biosocial view of Norm-based Social Learning. *Connection Science*, 18/3, 363-378. (2006).

[15] Cowley, S.J., Langford, D. & Schulz, J. Content coding: tracking a social mediator's achievements. Paper to be presented at *Human-Robot Interaction*, Amsterdam (2008).

[16] Craik, K. *The Nature of Explanation*. Cambridge: Cambridge University Press. (1943).

[17] Crespi, B. & Badcock, C. (in press). Psychosis and autism as Diametrical Disorders of the Social Brain. Behavioral and Brain Sciences (to appear).

[18] Dautenhahn, K. Embodiment and Interaction in Socially Intelligent Life-like Agents. In C. Nehaniv (ed) *Computation for Metaphors, Analogy and Agents.* Springer Lecture Notes in Artificial Intelligence, Volume 1562, pp. 102-142. (1999).

[19] Davies, P. S. *Norms of Nature: Naturalism and the Nature of Functions,* Cambridge MA: MIT Press. (2001).

[20] Davies, P. S. What kind of Agent are we? A Naturalistic Framework for the study of Human Agency. In: *Distributed Cognition and the Will*, D. Ross, D. Spurrett, H. Kinkaid & L.G. Stephens (Eds.) MIT Press: Cambridge MA, pp.39-60. (2007).

[21] Dennett, D. C. The Origins of Selves. *Cogito, 3,* 163-173. (1989).

[22] Dennett, D. C. *Consciousness Explained,* Boston: Little, Brown & Company. (1991).

[23] De Jaegher, H & Di Paolo, E. Participatory Sense-making: An Enactive Approach to Social Cognition. *Phenomenology and the Cognitive Sciences*, 6(4), 485-507. (2007).

[24] Di Paolo, E., Rohde, M., Izuka, H. Sensitivity to Social Contingency or Stability of Interaction? Modelling the Dynamics of Perceptual Crossing. To appear, *New Ideas in Psychology.* (2008)

[25] Fong, T., Nourbakhsh I., & Dautenhahn, K. A Survey of Socially Interactive Robots. *Robotics and Autonomous Systems* 42(3-4), 143-166. (2003).

[26] Gibson, J. J. The Implications of Learning Theory for Social Psychology. In: Experiments in Social Process: A Symposium on Social Psychology, J. G. Miller (Ed.) McGraw-Hill: New York, pp. 149-167. (1950).

[27] Goffman, E. The Interaction Order, *American Sociological Review,* 48, pp. 1-17. (1983).

[28] Hodges, B. Good Prospects: Ecological and Social Perspectives on Conforming, Creating and Caring in Conversation. *Lanaguage Sciuences,* 29: 584-604. (2007).

[29] Hutchins, E. *Cognition in the Wild.* Cambridge, MA: MIT Press. . (1995).

[30] Järvilehto, T. The Theory of the Organism-Environment System 1. Description of the Theory. *Integrative Behavioural and Physiological Science*, 33/4: 321-344. (1998).

[31] Kanda, T. Sato, R. Saiwaki, N. & Ishiguro, H. A Two-month field Trial in an Elementary School for Long-term Human-robot Interaction. *IEEE Transactions on Robotics (Special Issue on Human-Robot Interaction)*, 23(5), pp. 962-971. (2007).

[32] Legerstee, M. *Infants' sense of People: Precursors to a Theory of Mind.* Cambridge University Press, Cambridge.. (2005).

[33] MacDorman, K. F. & Ishiguro, H. "The Uncanny Advantage of using Androids in Social and Cognitive Science Research", *Interaction Studies, 7*(3) 143-158. (2006).

[34] MacDorman, K. Life after the Symbol-system Metaphor. *Interaction Studies 8.1: 143-158.* (2007).

[35] MacDorman, K. F., Minato, T., Shimada, M., Itakura, S., Cowley, S. J. & Ishiguro, H. Assessing Human Likeness by Eye Contact in an Android Testbed. *Proceedings of the XXVII Annual Meeting of the Cognitive Science Society*. (2005).

[36] Menary, R. Writing as Thinking. *Language Sciences*, 29; 621-633. (2007).

[37] Menary, R. *Cognitive Integration: Mind and Cognition Unbounded.* Palgrave Macmillan. (2007).
[38] Nabe, S., Cowley, S.J., Kanda, T. Ishiguro, H. Iraki, K. & Nargita, N. Robots and Social Mediators: Coding for Engineers. *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication. University of Hertfordshire*, pp. 384-390. (2006).
[39] Pattee, H. The Physics of Symbols: Bridging the Epistemic Cut. *Biosystems*, 60: 5-21. (2001).
[40] Ross, D. & Dumouchel, P. Emotions as Strategic Signals. *Rationality and Society, 16*(3), 251-286. (2004).
[41] Sommerhoff, G. & MacDorman, K. F. (1994). An Account of Consciousness in Physical and Functional terms: A Target for Research in the Neurosciences. *Integrative Physiological and Behavioral Science, 29*(2), 151-181.
[42] Tanaka, F. Cicourel, A. & Movellan, J.R. Socialization between Toddlers and Robots at an Early Childhood Education Center. *Proceedings of the National Association of Sciences*. (to appear).
[43] Thompson, E. *Mind and Life: Biology, Phenomenology and the Sciences of Mind*. Cambridge MA: Harvard University Press. (2007).
[44] Trevarthen, C. The Concept and Foundations of infant Intersubjectivity. In: *Intersubjective Communication in Early Ontogeny*. Cambridge: Cambridge University Press S. Bråten (Ed.), pp. 15-46. . (1998).
[45] Trevarthen, C. & Aitken, K. J. Infant Intersubjectivity: Research, Theory and Clinical Applications. *Journal of Child Psychology and Psychiatry*, 42(1), 3-48. (2001).
[46] Turing, A.M. Computing Machinery and Intelligence", *Mind,* 49, pp. 433-460. (1950)."
[47] Turkle, S., Taggart, W., Kidd, K. and Daste, C. Relational Artifacts with Children and Elders: the Complexities of Cybercompanionship. *Connection Science*, 18/4: 347-362. (2006).
[48] Vygotsky , L. The Instrumental Method in Psychology. In*: The Concept of Activity in Soviet Psychology*, J.V. Werstch (Ed.) Armonk, NY: N.E. Sharpe, pp.134-143. (1981).
[49] Walmsley J. Methodological Situatedness; or, DEEDS worth Doing and Pursuing. *Cognitive Systems Research,* (2008).
[50] Wenger, D. & Sparrow, B. The Puzzle of Coaction. In: *Distributed Cognition and the Will*, D. Ross, D. Spurrett, H. Kinkaid & L.G. Stephens (Eds.). MIT Press: Cambridge MA, pp.17-41. (2007).
[51] Werry, I., Dautenhahn, K. Ogden, B. Harwin, W. Can Social Interaction Skills be Taught be a Social Agent? The Role of Robotic Mediator in Autism Therapy. Springer Verlag, Lecture *Notes in Computer Science, subseries Lecture Notes in Artificial Intelligence.* (2001).
[52] Wheeler, M., *Reconstructing the Cognitive World: The Next Step.* MIT Press, Cambridge MA. (2005).

The Antiquarian Librarian & the Pedantic Semantic Web Programmer: Trust, logic, knowledge and inference

Cate Dowd [1]

Abstract. The logic and thinking for the Semantic Web environment has a philosophical base associated with rules and knowledge that draw on traditional concepts, such as 'loci' and 'indexes of place' for the location of items and information. These concepts and other characteristics that led to early library systems and were evident in print publishing, still have some links and relevance for the Semantic Web. For example, location and retrieval principles, once characterised by the 'loci' as a place for an idea, and the headings under which an argument could be found continue to inform the processes of mark-up, indexing and labelling for the Semantic Web.

Web programmers and librarians share important common tools and frameworks. In particular they are guided by sets of rules and logic that are used for the generation of information and knowledge. To generate specific knowledge, in particular for online learning, for example via an online epistemic game, a web developer must carefully design how, when and if information is accessed, which can be achieved via the use of algorithms. To design algorithms for game-play requires an understanding of various types of logic and the construction of values that are assigned to objects that need to translate into meaningful tools for a community of users. The values and decisions reached in the software design process will be informed by the appropriate identification of potential objects; the construction of classes; appropriate assignment of values; choice and selection for distributed data; the construction of ordered lists; labelling and so on.

Early design decisions for potential interaction can also be guided by a thorough understanding of the unique language and meanings that come from the communities of practice who will ultimately exchange information and generate knowledge as they use a system. Design approaches for complex applications, such as a serious online game requires an understanding of the need for multiple rules from which actions and interactions can occur. These rules and outcomes will be based on inference, which can draw on various kinds of logic.

Design features for serious interactive online environments also require structured approaches for understanding the lexical and hermeneutic base of likely end user participant communities. By examining the unique meaning(s) and *semantic* associations of a community's language and communication approaches, and by identifying and classifying potential game-play entities into a representational model, a programmer may gain rich insights to assist the articulation of entity relationships. From there it may be possible to build a formal Ontology which can become a base for an epistemic serious game in the web environment.

Various examples of mark-up languages over time, in particular for online news sites, provide a direct demonstration of the links between constructed language and communities of practice. A light analysis of various mark-up languages and some historic and philosophic perspectives on the construction of know ledge systems and knowledge management may well assist understanding of early software design issues for the Semantic Web. This is also a ripe time to at least raise a discussion about the nature of knowledge, which might emerge from a collective folksonomy, perhaps generated via online news tags for links with social sites, or might be distinct doctrinal or organisational knowledge generated via a formal Ontology, such as that needed for a serious learning environment.

1 INTRODUCTION

Many new forms of 'mark up' developed in the late 20th century based on the Extensible Mark up Language (XML), which changed dramatically the way people began to do business; generate news; and play, especially online and across continents. These applications generally addressed the needs and thinking of various communities of users.

Contemporary mark up languages did not begin with the advent of computing systems; rather they evolved from various non-digital systems, such as library systems which are characterised by indexes, labels and catalogues to assist the location and retrieval of information objects, such as books. The earliest kind of mark-up was a set of instructions within handwritten manuscripts to convey style and layout for the printer. The process of 'annotation was called marking up and the instructions themselves were often referred to as markup"[1]. The annotations related to 'presentation and structure', which are base characteristics of HTML documents, albeit for presentation and display output to a browser, rather than a printer.

Beyond the presentation and structural elements of web documents using HTML, the Semantic Web enables associative meaning to language through syntax, logic and XML server technology. The potential functionality of the Semantic Web depends on logic and information structures that can be understood through analysis of the activities and thinking of the librarian, antiquarian and/or contemporary.

To determine where a book might reside in a library system a librarian refers to a book index and works with a mark-up process to assist that decision. The title alone is too limited. The librarian, as an indexer, will decide the eventual location of the book based on a 'best fit' into a select subject area that matches

[1] Charles Sturt University, Australia, email: cdowd@csu.edu.au

the particular library classification system in use. The alpha-numeric details assigned to a book will make it a unique item within a whole system with potential for cross-referencing using a range of key terms.

It is the relationships and thinking at the cross-referencing level, seen in the library world that opens up the first-order logic associations that area also used in Semantic web applications. At the simplest level some item belongs to some other item. However, thinking and logic for the Semantic web extends to 'reasoning using probability and causality... and the relationships between statements, concepts, or propositions, and rules of inference'[2] .

Via the use of meta names or metadata tags, associations between objects can be made. The descriptive containers, whether for a library system or online news site become the base for building associations between objects that will be ordered according to imposed values and will be processed by levels of logic. An extension of simple associations between objects will build up to form a complex and structured algorithm and indeed artificial intelligence (AI) will emerge as sets of rules lead to knowledge-based outcomes.

In the early 21st century mark up languages express artificial forms of thinking and logic by defining sets of knowledge through entity relationships and corresponding rules. A Semantic Web application will be dependent on many logical features, but above all the design will require an understanding of the language and likely meaning(s) from the communities who will use it, whether the purpose is simply for meaningful search capabilities or more precise vocabularies for shared learning outcomes.

Online games and online news sites depend on precise vocabularies and classifications for objects, as well as rules, so that they can be manipulated by players and readers, or so that information can be shared via syndication sites. The Semantic Web and the exchange of document types now enables new associations and references for multimedia objects which has changed location and retrieval concepts. Indeed these concepts have now shifted towards the fast manipulation of objects, once only seen in stand alone applications.

For many communities the potential of logic via the Semantic web is only just emerging and it coincides with rises in computational speeds and powerful dedicated XML servers which can enable change for organisational process. However, the success of building Semantic Web applications and generating new knowledge requires careful discernments about the complexity of language in the construction of 'knowledge'. For example, multidisciplinary design teams need to come to grips with double hermeneutic situations that can arise from language ambiguity firstly within their own domains and then from a particular community who may use a Semantic Web application.

This paper shows how some linked characteristics of language, logic and thinking have evolved and transformed communication tools at different historic moments into potential knowledge. At the centre of this process is the art and science of indexing, labelling and mark up, and associated characteristics, some of which have remained the same over time. In a trajectory of technological change over time, significant attitudes, methods and tools associated with the location, retrieval and manipulation of objects have contributed to new knowledge and change for individuals and organisations. This paper identifies significant changes in thinking and the emergence of new ideas and structures for communication tools and systems over time, which still hold relevance for the design of knowledge management systems, whether for online news or online distributed serious game play environments.

2 DISTRUST AND THAT PLACE IN THE MIND (LOCI): PRINT TO SEMANTIC WEB

The early transitions from hand written scripts to printed documents during the 15th and 16th were marked by new society attitudes, in particular to the way writing and texts should be approached and used. These approaches contribute to new forms of indexing, mark up and labelling. Ironically, the changes that emerged came out of a mediaeval mindset based on a *mistrust* of the written word. The attitudes and responses to the printed word at the time have some traits that can still be identified in contemporary attitudes towards recent applications, such as the Internet:

> …to perceive writing as more than a reproduction of proximate speech required a leap of faith …they had to convince themselves that texts could be self-contained objects in their own right…and so they started to invent content devices for texts. These devices included signatures, seals, dates, locales, tables of contents, indices, and abstracts. The importance of these devices was to organise ideas according to the logic of the text (and the needs of the reader of the text) [3].

A simple comparison of early attitudes with contemporary approaches to the Semantic Web is most obvious in the areas of 'trust' and 'authentication'. However, the traditional devices also have numerous abstract concepts that still have significance for the Semantic web, such as logical associations between documents, location of documents, signatures and secure systems for transactions, to mention a few. Indeed *trust* and *signatures* not only carry on into the Semantic Web environment with similarities from earlier systems and times but are positioned on the 'top layer of the Semantic Web model'[4] whilst core technologies remain at a lower level.

Computing technologies use 'pointers' of various kinds, often for the location of a particular resource, or for the assignment of some value to an object. In the early print world a pointer may have been to a particular locale, a set of ideas within a written text or to a subject heading. In computing a pointer will point to a resource via a pathway to a particular resource, indeed even by typing a URL (a uniform resource locator) into a web browser.

3 INDEXES, LOCATION & COMMON PLACES

The concepts of location and retrieval have always been closely linked to indexing. To understand the diverse attributes of an index is more complex than first appears. For example, amongst the varies types of indexes used in libraries, even the earliest characteristics of the alphabetic index provide insight into the origins of 'relational' thinking that is still necessary for

the Semantic Web. Important characteristics also stem from the art of rhetoric:

> The alphabetic index is actually a crossroads between auditory and visualist cultures. 'Index' is a shortened form of the original *index locorum* or *index locorum communium,* 'index of places' or 'index of common places'. Rhetoric has provided the various loci or 'places; - headings, we would style them – under which various 'arguments' could be found, headings such as cause, effect, related things, unlike things, and so on.[5]

The idea of the 'argument' from oral cultures is no doubt less precise than most indexing in the 21st century. However, the notion of a 'loci' even if originally "thought of as, vaguely, 'places' in the mind where ideas were stored"[5] translates to an abstract location, even with empty containers, which are well reflected in meta name tags or a class of objects.

In the printed book, in mediaeval times "the vague psychic 'places' became quite physically and visibly localised."[5] The organised spaces and place for categorised texts within books were perhaps the first signs of refined notions for categories, indexical thinking and labelling. They offered structure for the location and retrieval of information and content within a book, which no doubt was more refined than simply pointing to a stack of books.

4 FIRST-ORDER PREDICATE LOGIC IN LIBRARY SYTEMS AND THE RISE OF FORMAL LOGIC

In the English speaking world a uniform approach to cataloguing and indexing emerged in the 19th century with the Decimal Dewey System. The classification system was, and still is, based on decimal numbers and subject areas with additional information about the time of publishing and other details. It was designed for ease of sorting, labelling and retrieving books and documents, and the potential for cross referencing within the system marked the beginning of a structured way in which to define simple sets of knowledge within a system, towards an information system.

The importance of relationships between resources, including categories for subject, date and type of document, brought about an entirely new system that by the late 20th century would be reflected in various digital environments, including the web. In the digital library environment resources would be labelled and sorted by the support of the Resource Description Framework[1], commonly called the RDF, which was integrated other library conventions, such as the Dublin Core, which defines the core descriptive terms for a library resource.

The evolution of logic to assist the location of objects or information artefacts using technological systems are characterised by the ongoing relevance of first-order logic, which is a form of predicate logic that indicates that a 'resource has a *property* and a value for that object'[2]. These features form the base of the RDF and can be interpreted within the library system as 'the book [subject] has the title [predicate] with a specific value [object]'. This formal type of logic has some similarities with other types of logic that have been developed within computing such as the 'Hoare triple, consisting of an assertion, a precondition and postcondition'[2]. The various forms of logic and the formal use of language have increasing relevance as the complexity of a system is revealed. For example a serious online game environment is characterised by many 'pre-conditions and post-conditions' that are essential, simply for game-play.

[1] The Resource Description Framework provides guidelines for specific and universal terms or tags in which to describe a type of content or attribute, such as geographic location of a resource.

5 FROM DISTRUST AND SETTING STANDARDS TO ORDER, LOGIC AND A.I.

By the early 20th century industrialisation had shaped printing, publishing and libraries and generated new attitudes and organisational processes. In particular, 'standardisation' influenced attitudes and process in areas such as 'standardised typefaces and typesetting conventions'[3]. By the 21st century 'the evolutionary appropriation of printing technology led to the construction of a communication artefact with the ... features of standardisation and mass reproducibility – an artefact whose widespread adoption has been associated with such major transformations as the coming of the nation state and the rise of modern science'[6].

Just as mass communication processes and artefacts have been linked to major societal change, the new logic and thinking associated with the Semantic Web has begun to transform our understanding of knowledge and knowledge management, and changed the ways in which we exchange information, interact and learn. However, the traditional tools and methods associated with indexing, labelling, mark-up and logic have increasing importance and continue to be refined in the 21st century, alongside ongoing principles for location and retrieval of objects, but the speed of location and retrieval is mostly invisible.

In addition more sophisticated forms of logic have emerged to achieve deeper levels of intelligence, including artificial intelligence, characterised by 'Higher Order Logic and inference rules'[2]. The relationship between artificial intelligence and knowledge is a large topic, but at a fundamental level it opens up an important topic for early software design in so far as *Tacit* knowledge from a community of potential users of a system must be transformed into *explicit* knowledge before any rules or inferences can be applied, and this can only be achieved once all manner of 'things' have the right labels.

6 [MARKUP] SHIPPING NEWS TO THE WORLD WIDE WEB

The early process of annotation for manuscripts was called *marking up* and the instructions themselves were often referred to as *markup*"[1]. Two early types of mark up for printers included "procedural markup for a typesetter with instructions about how to lay out text, for example, insertion of bold and italic type and different sizes, and descriptive mark-up which would indicate the type of content – for example, emphasis and chapter heading.."[7].

Mark-up for printed material in the 19th century was often very limited or virtually absent, especially for printed newspapers. For example, in Australia, any notion of style for presentation and layout in newspapers was extremely limited and 'until the

1860's…news simply consisted of paragraphs with no attempt at display'[8]. The attitudes and approaches at that time reflected the mindset and culture of the time. They were colonial times with very practical needs to communicate real world information whether for shipping news or parliamentary reports. They were not news stories as we understand them today, and the community of users would have had very different expectations compared to today's newsreaders.

The introduction of pictorial elements into newspapers and the use of mark-up for that end is a story of its own, but the development of mark-up shares some organisational roots that are noteworthy. Online newspapers in the late 20th century are guided by various international standards and protocols set by the telecommunications sector and the W3C[2] organisation, in particular for mark up and conformance, firstly to enable interoperability and to standardise the presentation and layout of documents. Presentation and layout standards in the web environment continue to share core concepts for mark-up with newspaper layout, typically for columns, tables and images.

These characteristics from newspapers, shared in the web environment have cyclical roots in so far as web standards are linked to telecommunications and the "the first newspaper in America…was published by the postmaster"[6]. The expert layout and mark-up knowledge may not have come full swing, but the link with postal and telecommunications concepts continue in the 21st century.

Today, the layout and *presentation* of online digital news formats requires strict vocabularies and rules that allow for far more *flexibility* and *style* than what could have been achieved with earlier printed documents. However, the core principles of layout, indexing, mark-up and labelling continue into the 21st century. Perhaps a small community of users are still interested in shipping news, but even they are most likely to seek information from a different kind of dock, and in real-time. Most of us are no longer waiting for our ship to come in - and we generally don't wait months to read about it, even if we are waiting.

7 CLASSIFICATIONS, METADATA & THE LABELLING OF OBJECTS

In the mid 1990's electronic library cataloguing systems in many countries migrated to web based interfaces. The importance of access to online information about printed resources led to a new set of standards to define the existing data related to collections, subject areas, location and other details for books and journals.

A new standard for 'metadata' called The Dublin Core was introduced as an ongoing set of initiatives from a working group for improved descriptions in the online environment using specific terms, labels and attributes for identification and location of library resources.[9]. The standards have gradually been integrated with other web standards for multimedia objects, for instance a video may be "assigned a DOI (Digital Object Identifier) number so that it can always be found, for intellectual property reasons or if its location on the web changes".[7].

The DOI for a chunk of video is just one example of a label that can be used in the digital environment. From a librarian's perspective the video might also have some 'faceted' associations that are determined by a controlled vocabulary [7], i.e. the object has a relationship to other related objects, which are defined by the metadata name.

The idea of objects having abstract associations is not new as is evident in Ong's discussion of how print was first regarded as an object that soon invited a label:

> Once print had been fairly well interiorised, a book was sensed as a kind of object which 'contained' information, scientific, fictional or other, rather than, as earlier, a recorded utterance…Each individual book in a printed edition was physically the same as another, an identical object, as manuscripts were not….with print, two copies of a given work did not merely say the same thing, they were duplicates of one another as objects.[5]

Steinberg, cited in Ong adds that "The situation invited the use of labels, and the printed book, being a lettered object, naturally took a lettered label, the title page"[5].

Alphabetic classifications or catalogues are visibly evident in the physical library shelves at Trinity College library, Dublin (See figure 1) where the letters of the alphabet can be seen on each shelf level. This simple alphabetic catalogue system represents just one type of cataloguing prior to the introduction of the Dewey Decimal System.

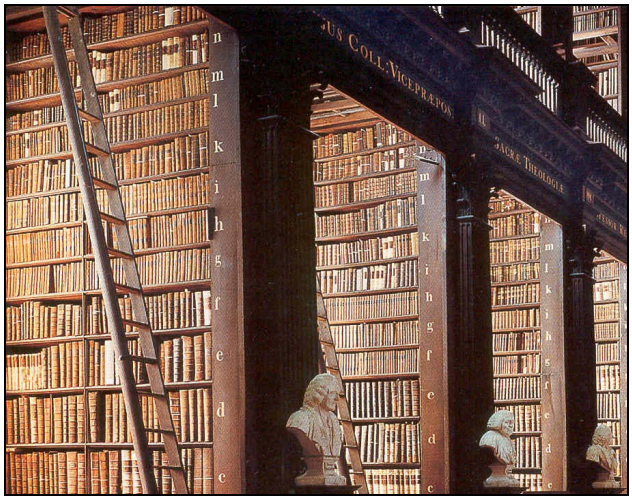

Figure 1. The book shelves of Trinity College Library, Dublin, with obvious signs of an alphabetic classification system.

In the late 20th century mark up in the digital environment was associated with a document type of definition (DTD) of SGML[3] for the basic display of information in a web browser, rather than for any display in a printed book. The type of document was called a HTML document, which was, and still is, just one type of document. Document Type Definitions are not limited to the

[2] W3C is the acronym for the World Wide Web consortium, consisting of various working groups who develop standards for the web to ensure interoperability, accessibility and levels of conformance.

[3] SGML is the acronym for Standardised Generalised Mark up Language, a Mark up language developed prior to the advent of the Internet.

web, for example book publishing has its own DTD[1], but HTML as a document type was limited in what it could do.

8 THE EXTENSIBLE MARKUP LANGUAGE (XML) AND THE RESOURCE DESCRIPTION FRAMEWORK (RDF)

Real-time sharing of information and real-time transactions via the Internet using the XML began in 1996. XML could describe storage layout and logical structures[10], which when combined with XML servers and new organisational processes, would change the ways that people did business, generated news and played online and across continents. XML took little time to trigger new perceptions of communication and the closing of distances. It also raised many issues associated with knowledge, such as whether information was from an authentic source and could be trusted.

The relationship between library systems and the introduction of XML should not be understated. The Dublin Core, mentioned earlier, as a base digital system for classifications complimented by the Resource Description Framework (RDF) for integration of library resources not only provided new levels of access and exchange of digital data for libraries, but has become the framework for other innovations.

The RDF is a framework for 'resources' and is concerned with logical associations between resources, which might indeed be objects. It has become 'a model of statements...about resources, and associated URIs [Uniform Resource Identifiers] that have a uniform structure of three parts: subject, predicate and object." [2] The logic and structured associations based on syntax and semantics associated with the RDF are the base of the Semantic Web. Whilst the RDF framework might be considered relatively simple it has relevance for the construction of a Semantic Web Ontology[4] which is necessary for the artificial intelligence of a complex online game.

For the modern, rather than the antiquarian librarian, the RDF is also still significant because it has become a model for the "expression of semantic information (meaning) [and can] assist interoperability of data, computer-understandable semantics for metadata, and better precision in resource discovery than can be achieved with full text search."[7].

9 INTEGRATED XML APPLICATIONS

Metadata standards in the early 21st century are made up of many works in progress, partly due to the opportunity for developers to build their own language or tags using the XML standards. An important multimedia XML-based mark-up language that has been in use since the late 1990s is SMIL, or the Synchronised Multimedia Integration Language, which is for 'integrating streaming audio and video with images and text for the web"[7]. The SMIL descriptions for integrated media forms mark an important shift from mark up that simply displayed static objects in a web browser. These descriptions changed the concept of mark-up for *presentation* and display within a web browser to a range of new ideas about the manipulation of data as objects.

Explorations to also describe even higher level detail for video content as new faceted objects have been in progress for several years. In particular via the standards associated with the Moving Pictures Expert Group (MPEG), such as MPEG-21 and MPEG-7, which are concerned with segmentation classifications for video attributes. Some levels of research have also steered towards "multi-level video indexing approaches using the RDF to contain both Dublin Core and MPEG-7 descriptions of the same content"[7]. These more sophisticated levels of integrated research can be applied to interactive video projects and digital television, but they also use levels of indexing and labelling that require serious modelling due to complexity and the need for representation across the design process, and they raise various issues for implementation beyond the scope of this paper.

Another example of integrated web technologies that epitomises the use of mark up, indexing and labelling, even in name, is Annodex. Annodex is a 'file format for annotating (indexing) time–continuous bit streams so that people can use text queries to search for video clips, then hyperlink to other video, audio or web content….it [uses] the Continuous Media Markup Language (CMML)' [7]. The idea of continuous mark-up perhaps captures the essence of XML technology, which seems limitless.

The various examples of mark-up clearly indicate how XML applications have drawn heavily on the core concepts of library systems, such as indexing, and various formal logic. The guidelines for implementing the Dublin Core in XML [11] are openly available to developers and might suggest that the range of mark-up applications will continue to grow and change the knowledge bases of today. The range of XML applications in recent years that are already pervasive include the use of 'SportsML and NewsML, and [perhaps less pervasive] Bookmark Exchange language (XBEL)"[12]. The use of XML in online news is perhaps the most dominant.

10 XML SCHEMAS AND THE INTEGRATION OF THE DUBLIN CORE FOR ONLINE NEWS

The RDF as an XML application makes it possible to define sets of knowledge through entity relationships, using some higher level logic. Contemporary mediated communication forms that have integrated these standards have been able to build new associative entity relationships, using shared lists and various indexes. In particular these can be seen within online news publishing and in syndicated news stories.

Integrated XML applications, such as the integration of the Dublin Core and XML schemas[5], can be easily identified in the source code of an online news site. In particular the links between document types are evident. For example, the source code from the Australian Broadcasting Corporation (ABC) news website[13] (see figure 2) shows a relative link to the Dublin Core tags, represented by the repeated acronym 'DC', followed by a metadata category name for each title and descriptive category. Other specific terms of reference for document types defined by the ABC are also embedded in the code.

[4] A Semantic Web Ontology is more sophisticated than a set of classes and objects because it also has a set of inference rules.

[5] Schemas are essentially different classes of documents with elements and attributes that conform to a document type.

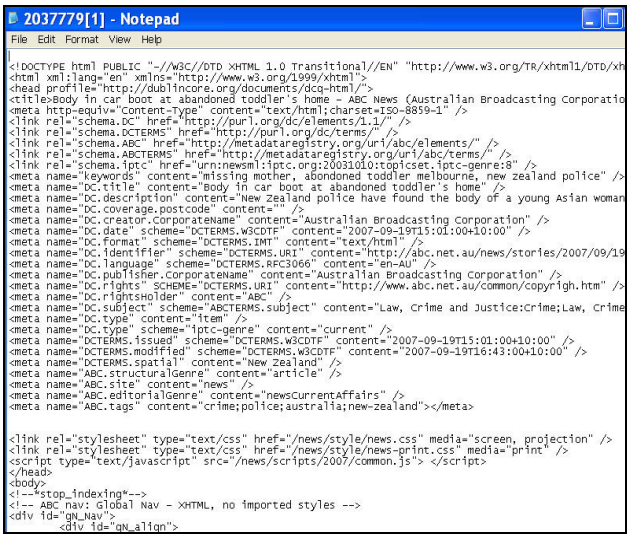

Figure 2. The source code from an ABC web site showing the Dublin Core terms and ABC terms to enable a resource, such as an article in the "news" category to be displayed in a Web browser. Source: ABC.

11 REALLY SIMPLE SYNDICATION (RSS)

The syndication of news stories is achieved using "RSS (Rich Site Summary) technology with metadata tags, which is XML based"[2]. RSS has actually evolved as three different versions and is more commonly referred to as Really Simple Syndication, or even news feeds, because it enables regular 'feeds' into an existing web page if a user chooses to make the appropriate links.

Digital 'tags' created by individuals or a particular community are the 'mark up labels' of the digital environment that have already generated some social change for mediated communications. The creation of shared tags is sometimes referred to as a form of "collective indexing"[7] and has facilitated the growth of Social Sites. Rather than using a formal taxonomy or even an ontology to define the nature of shared documents, social sites simply link to each other by agreement between folks, hence social sites are referred to in terms of "collaborative tagging and folksonomies" [7].

Current examples of social sites include Delicious, Flickr and Digg which provide digital feeds for photographic resources and news stories from a range of news sources, both independent and mainstream.

12 ONLINE EPISTEMIC GAMES: THE NEED FOR RULES, ALGORITHMS AND MODELS

An online epistemic game is generally designed as an immersive world for learning. The participant players do not simply scan web sites in search of higher levels of knowledge or undergoing traditional instructions from an avatar. Rather, they are situated in a constructed and simulated environment in which they learn about a professional role through engagement and immersion. Shaffer refers to this kind of learning or knowledge building as 'epistemic frames; collections of skills, knowledge, identities, values and epistemology that professionals use to think in innovative ways'[14].

The design of an epistemic game in the web environment requires deep knowledge of the participant communities and future players. In the early design phase it is crucial to represent and model the core areas for potential game play, which will slowly be shaped by the specific language and semantics of a community. It is also critical to understand the relationship between game play and rules, but also to make distinctions between doctrinal rules and inference rules which follow the limited logic of potential algorithms that might be developed. They are not the same thing. Nonetheless, in game-play 'playing a role means following some set of rules for behaviour'[14].

Online multiplayer games and simulations use various forms of mark up and scripting languages, for example Second Life as a virtual reality (VR) environment uses the Linden Scrip. This type of script defines many complex entities, including graphics and will process many computations to enable action. It draws on various forms of mark up, including XML and uses multiple forms of indexing, identification schemes and labels to ultimately enable the manipulation of many different objects, and at fast speeds.

A multiplayer environment can be modelled by 'classes' of objects, which are indeed the base of object orientated programming, but XML integrated applications will use schemas for classifying objects. There are many ways to begin to represent and communicate the entities of a future digital game play environment, such as using software tools for the Ontology Web Language (OWL)[6] or lexical based tools to help sort text and document types, like Leximancer[7]. However, the Universal Modelling Language (UML) (see figure 3) still provides an established base to begin building classes of objects and sub-classes that can represent and communicate core entity relationships and other features such as sequences of events.

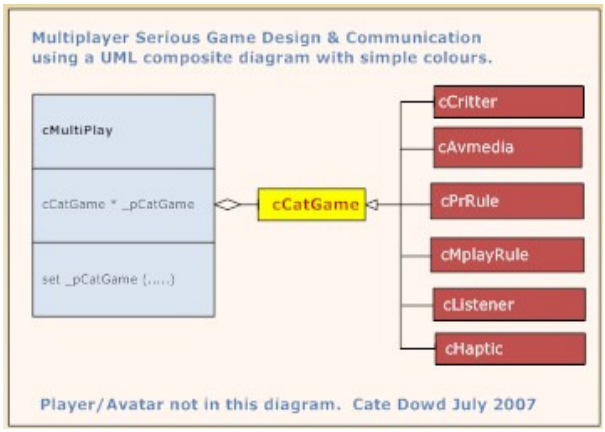

Figure 3. A UML diagram for a Serious Game design indicating potential classes of objects and a composite subclass with method and function.

[6] OWL is a recommended standard for processing content and information based on semantic meaning, that began in 2004 via the W3C http://www.w3.org/TR/owl-features/

[7] Leximancer is a software tool to assist text mining and is located at http://www.leximancer.com/cms/

The coded language, computations and processing in an online game can generate outcomes from player interactions via algorithms that will open up new choices for a player. The construction of multiple algorithms within a knowledge system suggests it is likely to be a complex system. However, for a system to achieve any levels of social intelligence, especially as a simulation, multidisciplinary design teams will need to grasp a large amount of information, which is perhaps best mapped via the use of formal models.

Models for representations of reality need to be simple and be able to communicate ideas during the design process. Object orientated analysis, towards design, aims to help 'represent ….real-world problem [s] in a format which a computer finds easy to deal with'[15]. As knowledge representations are developed via models they will inevitably begin to be embedded with various forms of logic that are necessary for an interactive system capable of actually managing knowledge. Ultimately the applied logic should assist developmental learning for a community of users if the system is designed as a serious game environment.

In addition to carefully planned learning outcomes, an online game-player will often experiences some sense of direct manipulation, which might be called an 'experience of agency'[16] or immersion. This could arise from navigation through a visual space with an Avatar or come from the effects of an explosion. In essence a script can trigger a momentary position and visual-auditory dimensions and associations between objects, which can be referred to as a pre-condition, which with a certain command, can generate a post-condition. The immersive experience is based on the mix of logic, graphic realism, and the rewards embedded in the design. From a design perspective, the game will have imposed values that aim to generate a learning outcome, in particular the design of a serious game. Clearly this will be a different shared experience than what might be generated collectively in a massive-multiplayer online game environment.

The hierarchical structures of the Semantic Web and the use of inference rules can also apply to a network of addresses drilling down to *locate* an individual machine responsible for particular computations in a distributed system, which may indeed be necessary for an online serious game shared by large professional communities.

13 CONCLUSION AND REFLECTIONS

The online game-player experience or the use of XML based RSS tags to allow a news reader access to syndicated online news stories, and even collaborative tagging, depend on multiple systems rich with indexes and labels. Some of the oldest principles of mark up, location and retrieval have clear pathways from print publishing and library systems into computing and information systems. They highlight the relevance of earlier efforts and the need for refinements, in particular for Semantic Web applications. The Resource Description Framework has formed an important base for higher level semantics alongside the use of multiple rules for inference and various forms of logic that will continue to be used for knowledge management purposes.

Information and knowledge management over time is clearly embedded in philosophical issues, in which human beings as creative developers must make absolute decisions about access and the construction of conditional situations for others. These decisions and situations are based on individual and collective value systems which can often be in conflict.

A socially intelligent computing system in the Semantic Web environment will be partly shaped by the applied logic, mark up, indexing and labelling associated with that system. Some of the characteristics of information and knowledge management systems discussed in this paper in various ways have been more visible in earlier information and knowledge systems and are now perhaps lost within the 'abstractions of a complex device such as a computer… [which many people still prefer to see] as a single, comprehensible unit'[17].

Perhaps more than ever the design and development of complex systems must look to deeper understanding and interpretation of content, in particular to the coded language and nuances of future communities of users, simply in order to get the game design right. This could be partly achieved through a formal focus on interpretation and the identification of 'hermeneutic' patterns across domains.

The location and retrieval of objects from within a socially intelligent system, digital or otherwise, is most likely to work well where objects or agents having unambiguous labels; are classified and grouped into appropriate indexes; conform to mark-up; and are subject to appropriate rules and various forms of applied logic.

The design of online news and more complex online systems, such a multiplayer game environment is a pedantic semantic situation for a programmer and a design team! It might be so, for several reasons, but it is no doubt partly due to the ambiguity of terms found across disciplines and domains and the need for very precise vocabularies in the Semantic Web environment. The tags and algorithms of the new artificial intelligence, like the catalogue labels, seals, signatures and stacks of the old, are still entwined in the topics of trust, distrust, logic, knowledge, rules and inference.

ACKNOWLEDGEMENTS This research is supported under the Australian Research Council's *Linkage Projects* funding scheme (project number LP0775418)

REFERENCES

[1] Chapman, N. and J. Chapman, *Digital Multimedia*. 2000: Wiley, New York.

[2] Alesso, P. and C.F. Smith, *Thinking on the Web: Berners-Lee, Gödel, and Turing*. 2006: John Wiley & Sons, Inc, New Jersey.

[3] Kaufer, D. and K. Carley, *Communication at a Distance: The influence of Print on Sociological Organisation and Change*. 1993: Lawrence Erlbaum Associates, Hillsdale, New Jersey.

[4] Antoniou, G. and F.V. Harmelen, *A Semantic Web Primer*. 2004: MIT Press, London.

[5] Ong, W., *Orality and Literacy* 1982: Routledge, London.

[6] Boczkowski, P., *Digitising the News: Innovation in Online Newspapers*. 2004: MIT Press, Massuchesetts.

[7] Brown, G. and J. Jermey, *The Indexing Companion* 2007 Cambridge University Press, Melbourne.

[8] Mayer, H., *The Press in Australia*. 1964: Landsdowne Press Pty Ltd, Melbourne.

[9] DCMI, U.B. *DCMI Metadata Terms*. 2006 [cited 24th September 2007]; Available from:http://dublincore.org

[10] W3C. *Extensible Markup Language (XML) 1.0 (Fourth Edition)*. 2006 [cited 22nd September 2007]; Available from:http://www.w3.org/TR/2006/REC-xml-20060816/#sec-origin-goals.

[11] Powell, A. and P. Johnston. *Guidelines for implementing Dublin Core in XML*. 2003 [cited 24th Sept 2007; Available from: http://dublincore.org/documents/dc-xml-guidelines

[12] XML.org. *Standards, Specifications and Initiatives* 2006 [cited 24th September]; Available from:http://publishing.xml.org/standards

[13] ABC. *ABC Online*. 2007 [cited 19th September 2007]; Available from:http://www.abc.net.au/.

[14] Shaffer, D.W., *How Computer Games Help Children Learn*. 2006: Palgrave Macmillan, New York.

[15] Rucker, R., *Software Engineering and Computer Games*. 2003: Pearson Education Limited, Essex UK.

[16] Murray, J., *From Game-Story to Cyberdrama*, in *First Person: New Media as Story, Performance and Game*, N. Wardrip-Fruin and P. Harrigan, Editors. 2004, Massachusetts Institute of Technology: Massachusetts.

[17] Brookshear, G.J., *Computer Science: An Overview*. 9th ed. 2007: Pearson Education, New York.

Could a Created Being ever be Creative? Some Philosophical Remarks on Creativity and AI Development

Y. J. Erden (Student)[1]

Abstract: This paper allots creativity a central role in enabling human beings to develop beyond the undertaking and/or fulfilment of simple primary functions. This contention is significant for Artificial Intelligence development since attempts to imbue artificially created-beings with ever greater levels of autonomy necessarily raises questions about the potential for creativity. This paper begins by highlighting key problems which occur as a result of attempts to offer definitive criteria for creativity, and uses this as a springboard to show that whilst certain AI programs may *appear* to present elements of creativity, the notion that these programmed acts could be creative in the sense in which we use and understand this term, is in fact mistaken. If we consider notions of creativity from a Wittgensteinian perspective we may yet achieve clarity about a subject which is often considered intangible and mysterious, whilst also coming to see its inherent irreplicability.

1 INTRODUCTION: A VAGUE DEFINITION?

It is of no small significance that to begin a paper on creativity requires creative thinking. A simple way to begin would be to identify possible definitions of creativity; however, such a task is made impossible by a general divergence of opinion not only on the content and structure of creativity, but even on its availability. For example, Martindale (1999) suggests it to be a *rare trait* because, he claims, 'it requires the simultaneous presence of a number of traits' (137). In a similar vein, Johnson-Laird (1988) maintains that inventions of new genres or paradigms are rare since '[t]here appear to be no common principles that account for such transitions within a field' (217). And even where creativity does occur, it seems the creative person may be in no position to offer any guidance on what took place since major innovations often depend 'on events of which the individual creators (and everyone else) is entirely ignorant' (217). If we follow Johnson-Laird's assessment, then a consequence of these facts for an enquiry into creativity (and the processes it follows) would be for there to be 'no general criteria or principles that underlie all and only the successful major transitions in a particular domain of art or science' (217). Ward et al (1999) offer a less pessimistic perspective however, and suggest that 'the capacity for creative thought is the rule rather than the exception in human cognitive functioning' (189). This divergence of opinion on the fundamental nature of creativity should not surprise us however, and I will argue in this paper, it is due to a common misunderstanding of the role terms such as creativity play in language. To explain this further it will prove useful to consider some typical definitions of creativity.

Common to most definitions is the assessment of creativity based on three specific criteria. Firstly, questions are often raised about the content and context of a creative idea or act, for example: is something creative *only if* it has been plucked out of the void (if that is even possible) or whether it is (could only be) the re-combination of a number of already existent elements? Secondly, questions are asked about the *value* of the creative product, for example, whether it is interesting or useful. Finally, weight is given to the question of intention and process: is the idea or act creative by virtue of what has been produced, by virtue of the process undertaken, or a combination of the two?

Martindale (1999) argues that the creative act involves the 'discovery of an analogy between two or more ideas or images previously thought to be unrelated. This discovery does not arise from logical reasoning but, rather, emerges as a sudden insight' (148). In addition to this, he notes creative inspiration 'occurs in a mental state where attention is defocused, thought is associative, and a large number of mental representations are simultaneously activated' (149). This is a definition which is shared to a greater or lesser degree by a great many commentators. Other definitions hold creativity to be 'the ability to produce work that is both novel (i.e. original, unexpected) and appropriate (i.e. useful, adaptive concerning task constraints)' (Sternberg, 1999: 3). Boden (1999) elaborates on this description,

> '[T]he generation of ideas that are both novel and valuable. *Ideas*, here, is intended in a very broad sense to include concepts, designs, theories, melodies, paintings, sculptures, and so on. The novelty may be defined with reference either to the previous ideas of the individual concerned or to the whole of human history.' (Boden 1999: 351)

Creativity has also been defined as 'the capacity for *variable focus*' (Gabora, 2002: 129), whereby 'creativity is associated with, not just high conceptual fluidity, nor just extraordinary control, but both'. In simple terms, association with the capacity for *self control* is as important for creativity as the capacity for remaining flexible. Each holding more or less favour as the situation requires. Other definitions of creativity include emphases on:

- Individuality, potential, personality, unpredictability, unique, surprising
- Interest, concern for, drive, judgement, motivation, cultural context

[1] Philosophy Dept., Roehampton University, SW15 5PU, UK, Email: j.erden@roehampton.ac.uk

- Originality, seeing things in new ways, freedom to be unbounded by convention and tradition, can change direction and approach, unfixed
- Inspiration, answering adversity, knowledge, relevant skills

But this list is by no means exhaustive. It is clear that trying to define this term is not getting us very far. Saunders (2002) highlights the problems associated with attempts toward definitions, remarking that

> The apparent need to define the nature of creativity has haunted most attempts to develop models and theories of the processes involved. The difficulty of this task is clear from the number of definitions that can be found in the literature – Taylor gives some 50 definitions. Some researchers have concluded that trying to develop a single definition of creativity is a fruitless task and have looked for ways to conduct their research without the need for a formal definition. (80)

This last point is an important one, and one I shall give further consideration to at the end of this paper. Part of the problem of defining a term like creativity is that (in everyday use) we might be unclear about what criteria applies in a given situation, yet in spite of this, still be able to use the term meaningfully. This aspect should not be overlooked in trying to get to the bottom of the creativity question. It is a direct consequence of this that we have been unable to offer a clear definition, and yet this failure is not to be attributed to faulty reasoning or a yet unsolved puzzle. Rather it shows something quite profound, namely that we are dealing with a concept that appears to have contingently vague elements at its core. Attempts to define the term conclusively serve only to highlight this insurmountable dilemma. Yet in spite of this linguistic difficulty, my paper begins with the premise that creativity (understood in simple terms) is the single most important aspect of our ability to develop beyond primary functions—it is the fire which fuels our potential—even if it remains one of the more elusive aspects of our being. What we need to understand therefore is not why it is important (since this seems rarely to be doubted), but why it should appear elusive.

I begin an answer to this question by first making clear that the accounts offered above share one common premise, namely that creativity can *in principle* be defined *in and of itself*. It is this factor above all others which has led to this sense of its meaning being elusive. This is in part due to its inherent complexity, affected as it is by multifarious factors. It is not that any of the terms used above to describe it are *wrong* or deficient, but rather that no single term on its own is either sufficient or definitive. The fluidity of terms with which we could describe creativity is therefore a reflection of its own status as a particularly fluid term, all of which results from its application in a multitude of different language-games.[2]

Neglecting these language-games, or contexts which give meaning to the term, and attempting to analyse it within theoretical mediums is doomed to fail. As Wittgenstein explains, often it is as simple as accepting that 'the meaning of a word is its use in the language' (2001: §43). Positions which define terms separate from use belie metaphysical ambitions. They assume we can adopt a position *outside* our habits and practices, aims and abilities, and from this position offer a meta-theory which encompasses all content and structure. As the lack of consensus above shows however, this does not work, and the reason I am suggesting for this is that it is simply impossible to capture the meaning of a word separate from an understanding of the language-game within which it has meaning. In the same way that one comes to understand what 'pain' is within the different contexts that the word is used (I have a pain *here*, or this causes pain, and so on) creativity *means*, and comes to have meaning, through its *use*. Thus there is no essence to capture, or that might be captured, in attempts that offer a singular definition. And if the above shows anything it is that 'clear cut' definitions are unlikely to account for all actualities. The same problem occurs when we try to define a word like 'game':

> How should we explain to someone what a game is? I imagine that we should describe *games* to him, and we might add: "This *and similar things* are called 'games'". And do we know any more about it ourselves? Is it only other people whom we cannot tell exactly what a game is?—But this is not ignorance. We do not know the boundaries because none have been drawn. To repeat, we can draw a boundary—for a special purpose. Does it take that to make the concept usable? Not at all! (Except for that special purpose.) No more than it took the definition: 1 pace = 75 cm. to make the measure of length 'one pace' usable. And if you want to say "But still, before that it wasn't an exact measure", then I reply: very well, it was an inexact one.—Though you still owe me a definition of exactness. (§69)

Substitute 'creativity' for game here and you have a fair description of what the theorising described above has led to. Just like a 'game', to be 'creative' means any number of things, in any number of contexts. This is not to say that definitions of creativity have no place in analysis. Indeed I have made much use of it in this work, and it is of particular use where it serves the purpose of explanation, or so that judgements can be made about claims to creativity. Rather it is the definition which claims to be comprehensive of which we should be suspicious, since it will necessarily exclude more than it includes, and thus offer us an impoverished account of things. The consequences of this argument for AI development and creativity are significant, but before I come to this there remains one further aspect of creativity which requires some consideration, and that is its status as something 'mysterious'.

2 CREATIVE MYSTERIES

Profound questions arising from aspects of what the creative process appears to entail often claim creativity to be arbitrary, and it may be partly due to this that the entire process has traditionally been construed as mysterious. The first point is one to which I shall return shortly, but before I do, it seems prudent to first consider the claim to mystery that creativity has been accorded. Materialist accounts of mind would hold creative processes to involve brain functions not dissimilar to any other,

[2] Language-game is a term which brings 'into prominence the fact that the *speaking* of language is part of an activity, or of a life-form' (Wittgenstein, 2001: §23)

whilst those with metaphysical tendencies would contend that it is in fact *inherently* mysterious, and might conclude from this that it is thus, by its very nature, fundamentally unreplicable.

Schank (1988) disavows any notion that creativity is *inherently* mysterious, exclaiming that to an AI researcher this simply means that 'there seems to be no algorithm behind the creative process', but 'that such an algorithm must exist, in principle' (220). Boden (2004) echoes similar sentiments, and in this vein she also offers an argument against the notion of *intuition*, in this particular case as playing a role in aesthetic judgement, which might prove illuminating to our discussion here:

> to say that we do something intuitively does not mean that some power of intuition is involved. It means, rather, that we do not know how we do it. 'Intuition' is the name of a question, not of an answer. Moreover, it is a question that can sometimes be answered with the help of computer models (1999: 362).

There are two contentions to raise here, the first concerning the actual processes, and the second concerning the terms themselves. On the first account, Boden's assessment would seem plausible if intuition were said to be *undiscoverable*. In that case her claim for this step being pre-emptive would stand. In fact, her claim is stronger than this, and belies tendencies in the scientific approach to assume that many folk psychological terms such as intuition or belief may be eventually explainable (seen also in Schank's comments above, and Churchland, 1988). The problem with this idea is that it relies on the notion that *all things* will be ultimately explainable, and follow a logic which is in some way comprehensible. I would not claim that these beliefs are mistaken *per se*, or that there will never be discoveries which might account for what we consider intuition to be, but rather that it might be as likely that we will find answers for such questions, as that there might also be none. Or, at least, none that would fit any standard conception of what an answer in strictly scientific terms might be. If, for example, it came to light that intuition is in fact a term made from a complex of components, any of which might alter without having any significant effect on the intuitive capacity itself, then it would seem difficult to answer precisely *what* intuition might be.

The second contention concerns the same issues raised in the analysis of creativity above. Simply put, intuition, like creativity, is a term which refers to concepts that have meaning in particular language-games. In which case, intuition *means* little more than having a sense of what something is or might be, seemingly with little or no explanation as to how, where or to what this sense might be attributed. As such, this term would appear to be *necessarily* vague, since to be anything else would be to change the meaning of the word. The same can also be said of creativity. On this account, if we could answer Boden's question we would necessarily be discussing something that had nothing to do with intuition. This is not to claim that we might never be wrong in our reference to 'intuition', but rather that when we are wrong this shows not that our use of the word *itself* was wrong but rather that we were mistaken in our application of it. In addition to this the compositional nature of ideas can make them difficult to chart in any *mathematical* way. On this Gabora (2002) asks:

> Is it possible to mathematically model the creative process? One big stumbling block is that a creative idea often possesses features which are said to be *emergent:* not true of the constituent ideas of which it was composed. For example, the concept *snowman* has as a feature or property 'carrot nose', though neither *snow* nor *man* does. (130)

In light of these arguments it therefore seems peculiar to search for the meaning of a word which already lies open to view. An investigation of this nature assumes that a meaning is lacking and can only be found empirically. Instead, a simple re-assessment of how these terms have meaning shows that they do in fact stand in good order precisely as they are. There is nothing mysterious about the word 'creative' when I say that someone or something shows creativity, for if there were, how would you ever understand what I mean? Trying to answer the question of what creativity or intuition is, without recourse to standard definitions—since these definitions are at best vague—implies that vague definitions are *ipso facto* in need of clarification. And yet, being vague or unaccountable seems to be significant in what these concepts mean. Accordingly we could then attribute the 'mysterious' element of such terms to little more than a misunderstanding of language, and attribute this mistake to our language being 'on holiday' (Wittgenstein, 2001: §38). In fact, the process of being creative might prove no more or less mysterious than any other given psychological term (what it is to *believe*, for example). What we must now consider is how we might, if indeed we can, set about replicating such vague concepts in AI programming.[3]

3 CREATIVITY AND AI DEVELOPMENT

AI development has struggled to replicate human creativity in its entirety. As already noted, this has been due in part to our own difficulty in understanding what this term might consist in, as well as a general inability to offer criteria for what should and what should not be called creative. These aspects prove significant for attempts to imbue AI with the potential for creativity, but before I consider this further I need first to analyse some perceptions concerning the *content* of creative ideas, and to do this I will make use of two categories of creativity offered by Boden (1999). She divides creativity into two types: exploratory (E-creativity) and transformational (T-creativity). From these there can also be a combination of the two, which she terms ET-creativity. She explains that in AI it has been easier to model E-creativity rather than T-creativity (353). One example she offers to support this claim is with the BACON program. She says,

> BACON and similar programs can find linear (and other) relationships between measurements. But they have a built-in expectation that such relationships may be there to be found. In the history of science, the mere idea of asking such a question was a very creative (very significant) step.' (Boden, 1999: 359)

As such, the creativity displayed is exploratory (undertaken within given perimeters), rather than transformational, which

[3] Whether replication is a useful tool in artificial creativity development is another question and one to which I return at the end of this paper.

would, on this account, have more claim for being unbounded. Another pertinent example she cites is with the AM program, a T-creative program working on mathematical problems which includes 'heuristics for altering concepts'. There are, however, problems with some of the creative aspects since 'AM has picked out an enormous number of ideas that human mathematicians regard as boring, or even valueless' (365). Furthermore, 'AM is unable to change its own values: its criteria of what is 'interesting' never vary' (Boden, 1999: 365). What seems to be missing in these programs is the potential to change tact and to *choose*, and it is to these aspects that I will turn in the proceeding section.[4] Before I do however, are some reservations about Boden's account which require some attention.

The central objection is that this account premises exploration and transformation to be the most significant defining aspects of creativity. As Novitz (1999) points out: 'not all radically creative acts involve deliberate attempts to transform conceptual spaces' (pp.71-2). To this comment I would add that it does not follow that those that are not T-creative are necessarily E-creative by default. Both these terms are themselves complex and beg further questions, besides which it is often the case that what we consider creative remains open to debate and alteration. Nowhere is this more apparent than in the second criteria I noted at the start of the paper concerning *value*.

Novitz (1999) suggests that value has a key role to play in whether we consider something to be creative or not. For him it is requisite that 'a creative act be of real value to some people', and further that a 'recombination that appears to be valuable, yet is later found to be thoroughly harmful and of no lasting benefit to anyone, will not be of real value and so will not be creative' (77-8). This claim is a complicated one, and while it offers some support to my claims here, it is also problematic. In making this claim Novitz applies normative conditions to notions of creativity which are not dissimilar to those advanced by this paper, yet, in the manner in which he achieves this, he effectively rules out any possibility of claiming anything to be creative beyond claims for its *current* status. It is clear (as already noted) that what we consider valuable or what we consider harmful are dynamic concepts, and often in their application we make judgement calls which may stand or fall at any given moment (what we think good for us today is as likely to be considered bad tomorrow as to remain a good). On his account, the application of the term 'creative' would seem to rely on an unknown future factor, and thus would remain perpetually uncertain. Simply, we could never consider something creative with any certainty on account of our not knowing whether it was a *definitive* good. While 'value' is an important tool in assessing claims to creativity, it is by no means certain that claims for creativity would always be rescinded following shifts in perception of value. It seems that attempts to pin creativity down have been thwarted once again, and while both the accounts offered here have merit, they both also suffer from the same restrictive tendencies. These objections will have significant impetus for the question raised in the title of this paper, and it is to this that I now turn.

The question whether computers could be considered creative over and above notions which hold them to simply replicate the originality of the programmer is a serious one, but interestingly it is often dismissed by programmers as having significance only within philosophical settings. Boden (1994) claims that since it is 'not a scientific question' it can be ignored, since it is 'in part a philosophical worry about 'meaning' and in part a disguised request for a moral-political decision' (85). The problem with this way of thinking is that even if the question is ignored, this does not remove the difficulties associated with affixing terms like 'creativity' to machines *without* first getting clear about what we mean by creativity. For example, Boden poses the question above in terms of 'genuine creativity' (1990, 286), but this belies an assumption that there might be 'false' creativity. As this paper has shown, what we consider creative depends on a multitude of competing factors, and the term is assigned on condition that certain criteria are met, as specified within particular language-games. What *genuine* creativity is within this is just what is *accepted* as creative under any such criteria. This is not to say that we might not be mistaken about whether something is or is not creative, nor that we might not change our minds, but rather that something either is or is not creative according to the use of this term in language, and that as such, to assert creativity is to make use of a term which holds for those language-users who share our games. In the same way that if a lion could speak, he would not speak *our language* (and I will explain what I mean by this a little later on), even if a machine could be imbued with creative abilities similar to our own, they would not be *ours*. The computer, like the lion, does not share our form of life, and therefore cannot share our language-games.

Before I move on to questions concerning arbitrariness in creativity I have one final reservation that needs addressing regarding AI and creativity, and this concerns the notion of rule-following in relation to programming. Programs rely on systems within which symbols are manipulated according to formal rules. These, in turn, encode a set of properties. The problem for the replication of creativity within such systems becomes apparent when we consider this process in relation to Wittgenstein's remarks on rule-following. He explains that a rule may not be understood separate from its context (*our* context), and from this we can surmise that codifying a rule would not therefore be possible *in advance of* the practice in which it applies. Rules adopted in practices or judgments of creativity are only signs, they don't tell us *which way to go*, thus:

> A rule stands there like a sign-post.—Does the sign-post leave no doubt open about the way I have to go? Does it shew which direction I am to take when I have passed it; whether along the road or the footpath or cross-country? (2001: §85)

Often a misunderstanding of the nature of rules has led to serious errors in how we come to understand the creative process, for example, Jonson-Laird (1988) states,

> It is often claimed that a creator 'breaks the rules' in order to produce a more original work of art, Likewise,

[4] Since what drives creativity encompasses a wide variety of values, it is likely that tensions between competing factors may sometimes occur. For instance, if creativity is aroused in order to fulfil a particular need, and where the fulfilment of that need is detrimental to some other equally important need, then decisions will need to be made concerning the claims to value of these two competing needs. This, in turn, may result in important decisions being taken. The generation of such conflict (where it occurs) might therefore prove an important aspect of creative thinking. As Johnson-Laird (1988) point out, key factors often 'cannot be foreseen at the time of the innovation' (217).

> although a grammar may capture a genre, individuals have their own unique styles. Both these objections are instructive, but not decisive. If a creative process breaks the rules, then either it must make a choice at random regardless of the consequences or it must be governed by yet further criteria. These criteria can in turn be captured in a grammar. Hence, the breaking of a rule can be described by yet another rule (or else it is merely an arbitrary infraction). If an individual has a unique style, then it must depend on idiosyncratic biases in choosing alternatives. A grammar can likewise be framed to capture this style. (212-3)

Although working along similar lines to those I advance in this paper, the above remark shows a key difference in the conclusion reached. As discussed above, a rule can be understood only within context, but furthermore, it is something that can in principle be explained (or defined), otherwise what is described as a rule would in fact be better considered a way of doing things. Equating a 'style' with a grammar or a rule implies that the style is more regulative than it need be. Often a style can diverge from the usual, and may offer no more evidence of its heritage than a few minor traits which belie the stamp of its creator. Furthermore, as already noted, rules cannot be accounted for *in advance* of the practice, but are in fact tied up within it. It seems problematic therefore to rely on rules in order to predict which way the style will go. Rules are merely signs after all, and do not tell us where our journey should lead us. In this respect the notion of family resemblance might prove more useful here than that of rule-following. Particularly when the search for 'rules' is likely to slow down attempts toward understanding what creativity is in practical terms, as well as broader notions of how creativity *works*.

The problem is that programs which attempt to account for creativity appear to do just this. The rules upon which they are founded would need to encode a never-ending list of associated rules, and furthermore, there would seem to be the need to encode rules which would exist only in order to break other rules that may come up, as the above remark by Jonson-Laird makes clear. To do this however means we would need to understand a rule in advance of its practice, since an element of prediction seems necessary if we are not simply to generate random data. In fact, even if creativity has elements which are rule-based, this would not necessarily mean these rules could be codified, attempts to do so rest on a misunderstanding of what a rule is. Claims (cited in the section above) that the random element in the creative process is irrelevant seem fair in the sense that, for example, we often use the rule that *intention* should be part of a creative act, yet even here we cannot say that such rules could be codified, not least because it is difficult to offer certainty about what that intention must consist of.[5] I will return to these issues in the final section of this paper, but for now what is clear is that these reservations regarding rules and attempts to define perimeters of what we would call creative, echo those issues raised in the preceding sections about definition and language-games. Rules have meaning only within the structure that they are applied and followed, and it seems unlikely therefore that a set of rules about what creativity is could be designed and implemented outside of this structure.

4 FREEDOM AND CREATION

Johnson-Laird (1988) suggests that there are certain necessary factors which have significant effects on creativity. He contends that there needs to be a certain amount of freedom, which in turn allows one to choose to make arbitrary decisions: 'creativity depends on arbitrary choices and thus on a mental device for producing, albeit imperfectly, nondeterminism'. What this means is that given the same situation, 'a genuine process of imagination could deliver a different response the second time around'. Freedom is key in this because one 'demonstrates freedom (if not imagination) in acting arbitrarily' (207). He further claims that '[b]ecause the creation of new genres and paradigms is so difficult, it might depend on an essentially arbitrary or random generative process' (217). The potential for making arbitrary decisions pose few problems for AI development, since this has proven to be easily replicable. As such I shall focus here on the notion of free-will in relation to such decisions, although it still remains to be seen how a term like 'arbitrary' factors into what creativity is.

Johnson-Laird's account implies that a certain amount of autonomy is requisite for creativity to be genuine. Despite this, many accounts of creativity also claim that significant portions of these processes occur at unconscious levels, aspects of which the creative person herself would be unable to account for or explain. The problem with this scenario for AI development is to try to evaluate where the balance of favour lies. Is creativity primarily a conscious or an unconscious act? And to what degree is either aspect contingent for true creativity to occur? If, for example, a composer claims to have awoken from a dream with a melody fully formed, and claims this in earnest, would we really want to suggest that she has been creative? Let us suppose that on waking, the composer merely copies that which she had heard in her dream, and more importantly, to this she makes *no alteration*. Can we truly say that what has occurred is *creativity*? We can also imagine a contrary process, whereby a poet is instructed to write a poem following very strict guidelines. He diligently sits at his paper and writes the first words that come to mind, with entirely arbitrary choices for which words he writes, and no structure to these words. He then cuts each word into a small strip and puts these into a hat. After mixing these around, he pulls each word out individually, and writes them on the paper in the order in which they come. Let us suppose (however far-fetched) that this jumble of words is included in a poetic anthology of more traditionally composed poetry, and that it achieves some acclaim. It is successful in poetry terms. Since some of the criteria for what creativity is often claimed to be has been met (novelty, value) would we therefore want to ascribe the term 'creative' to the poet, poem, or even to the process by which the final result was achieved? These are the questions we face when we consider the AARON program.

AARON, a series of programs for generating line drawings, and more recently also for colouring them, has had its 'aesthetically pleasing' works exhibited in the Tate and around the world. The creator, Harold Cohen, on being asked whether AARON was being creative in such work replies

[5] What intention might be begs further questions about aesthetic criteria, as well as broader questions about the application of psychological terms in matters of judgement, but in the interest of brevity, I shall not pursue these here.

> I think creativity is a relative term. Clearly the machine is being creative…to the degree that every time it does a drawing it does a drawing that nobody has ever seen before, including me. I don't think it's currently as creative as I am in writing the program. I think for a program to be fully creative, in a more complete sense creative, it has to be able to modify its own performance, and that's a very difficult problem.[6]

As Boden (1999) points out, 'AARON cannot reflect on its own productions, nor adjust them so as to make them better' (363), and this very crucial element is one that is often cited as evidence against the possibility of genuine AI creativity. And yet one might counter-claim that this is actually no different from the case of the composer who awakens from a dream, since she would appear to have no further *conscious* autonomy in the production of the piece of dream-music than AARON has in producing a work of art. Of course, the composer could choose not to ever write the music she has heard in her head, but this aspect does not affect the point being made here, if for no other reason than that this would be a question of free-will and autonomy more generally, and not one which has bearing on the creative act in and of itself.

It seems clear that the question which AARON's artistic powers provokes is not whether the machine has successfully created a work of art that is in some way *aesthetically pleasing*, but whether this is what it is to be creative in art. Thus when Boden discusses the different achievements of two separate genetic algorithms (GAs) in aesthetic terms, it seems the boundaries between these two aspects are problematically blurred. She states that many people see one AI model (Sims' GA) as *more creative* than another because 'it always comes up with at least some patterns they regard as attractive'. The comparison drawn is with those produced by a different algorithm—Latham's GA—primarily because these are stated to be 'strongly repellent' due to featuring images 'which resemble molluscs and snakes' (367). It seems that creativity is being (mistakenly) equated with aesthetic judgement or appreciation. An artist is not deemed to be such because of the quality of their work, but by the work that they undertake. Just as a terrible baker will still be a baker, so too a bad artist is yet still an artist. It therefore seems problematic to apply such terms in our judgement of whether or not an AI program is creative or not. Particularly since it would be difficult to say of the work of some rather acclaimed artists (Damien Hirst, Tracy Emin or Francis Bacon to name but a few) that the term 'attractive' forms any part of an aesthetic judgement of the work.

Perhaps we will be closer to understanding the reluctance to equate what might be termed AI-creativity with that of humans when we consider such undertakings in terms of *ambition* and its relation to free-will. As Boden explains, even when a Sims genetic algorithm gives the appearance of transformational creativity, because it can 'make random changes' (367), these are not *focused* attempts in the same way as those made by the creative artist or scientist (Boden, 1999). Clearly these aspects are further indications of an agent's freedom to follow their own creative urges, yet could it be argued that these aspects are insignificant psychological or social aspects of the creative process rather than contingent factors?[7] Although there are strong objections to this way of thinking, as I have already made clear and to which I shall return below, let us for the moment suppose that such claims are valid, and that all such social aspects should be dismissed as redundant. If we succeed in this it makes the claim that AI programs display 'choices' in so called 'creative' acts easier to abide. In drawing one thing not another could it be said that AARON has made a *choice*? But what other factors are important in the replication of choice?

It may transpire that replicating genuine creativity—as, for example, displayed by those successfully creative persons—is only a matter of developing programs which contain more information, are more complicated or contain a larger number of competing factors. Computer programming, as with most disciplines that follow scientific or mathematical processes, is accumulative, and therefore we might grant that it is at least *possible* that all such factors could be accounted for and codified some time in the future. Even if the creative capabilities of successive programs were to prove deficient in some way, this might prove no worse than has typically been the case in advancing the development of other sorts of ideas. The question might thus prove to be more a case of 'when', rather than 'if'.

But there is a serious flaw with this thinking since even if one could codify all that a person *knows* into a computer, it is unlikely that this artificial body of knowledge could then replicate all the different ways in which that person might connect what are sometimes (apparently) disparate pieces of information. The computer might make *better* connections, but this would still be different. This is not to say that all humans make the same connections, but rather that by sharing a common language and a form of life, we are apt to make similar sorts of connections, or at the very least be capable of understanding even those radically different connections that are made by others. Part of what it is to be creative is the ability to look at things from different angles, or as Wittgenstein suggests, to see something 'in *this* way or *that*' (2001: §74). While there may appear to be no particular, or at least no over-arching, reason for why we see things one way or another this may belie a multitude of different or competing factors. It follows from this reasoning that the earlier dismissal of so-called psychological or social factors was seriously misguided. We simply cannot get away from the fact that our creativity is shaped by the very particular ways in which we come to see things in this aspect or another.

[6] Comment taken from the film *The Age of Intelligent Machines* by Ray Kurzweil (1987)

[7] This is also the case with respect to the urge toward creativity as response to feelings of restriction. What it is to be free is clearly measured in degrees, and it seems difficult to ascribe this to the production of creativity since doing so begs questions of how much freedom is needed in such cases. It is sometimes the case that we might be at our most creative in times of adversity and against opposition. Whether we *feel* free or not is also a significant factor in such questions (as Sartre acknowledged in his later existentialist claims). Further, as pointed out by my colleague in discussion, were one to have utter freedom, say from death in the form of eternal life, would there ever be the *need* to be creative? From this example, as with many others I have raised in this paper, we come to see that in trying to define the creative process, we often unwittingly limit what must by definition be unbounded, at least in potential, if not in reality. It is perhaps this that is most difficult to replicate. With this comes a certain amount of self-awareness and knowledge, but also a certain amount of optimism and drive. Futility is the one obvious destroyer of creativity. In all other respects, constraint may often prove just as stimulating to creative impulses as the freedom to choose.

And the reasons why we see things one way or another, or even how aspects just *do* appear to us, simply could not be codified, there would be too many variables, as this remark by Wittgenstein draws out:

> The concept of 'seeing' makes a tangled impression. Well, it is tangled.—I look at the landscape, my gaze ranges over it, I see all sorts of distinct and indistinct movements; *this* impresses itself sharply on me, *that* is quite hazy. After all, how completely ragged what we see can appear! And now look at all that can be meant by 'description of what is seen'.—But this just is what is called description of what is seen. There is not *one genuine* proper case of such description—the rest being just vague, something which awaits clarification, or which must just be swept aside as rubbish. (2001: pp.170-1)

Vagueness is thus as important here for the content of seeing, as it is to the definition of creativity. One final point about the body of knowledge claim before I move on. Langley and Jones (1988) suggest that there is a body of knowledge from which creative people create, and this would seem to support the argument I offered above for how creative choices might be replicated in AI. They explain:

> We have seen the important role that preparation plays in scientific insight, and presumably any creative act must have substantial knowledge structures on which to build. One cannot expect to be creative in any domain until one has achieved knowledge of that domain. (199)

There are two concerns which arise from this way of thinking however. The first is that it seems to suggest that all creative processes take place within the accumulative method of the sciences. This is clearly not the case, and although it is the most common form of paradigm change, it is not by any means the *only* form. This point leads on to the next, which is that it may well prove valuable (in certain situations) to look at something with fresh eyes, to not be burdened with tradition and how things *should* be. In fact, academia and vast quantities of knowledge can sometimes prove restrictive to creative thinking. Novitz (1999) makes the same point when he states (in response to Boden), 'it just is not true that radically creative human beings must always have explored and will always be familiar with the conceptual spaces that their ideas transform', since sometimes 'the weight of those domains, the pressure of orthodoxy, prevent them from noticing new possibilities, new ways of doing and conceiving' (71-2). In this respect, as in many others which I mention here, psychological factors may prove as important to creativity as brain activity (if the two can be divided thus). Thus, the body of knowledge and the problem of choice remain significant ones in creativity and AI development.

5 CONCLUSION: IF A MACHINE COULD *SPEAK OUR LANGUAGE…*

In *Philosophical Investigations* Wittgenstein enigmatically remarks: 'If a lion could talk, we could not understand him' (2001:190). In offering some answers to what Wittgenstein means here, I will show how his comment offers potential answers to the problems that arise when programming AI to be creative. Wittgenstein's remark is situated in a work which seeks to show how language is *used*, and in so doing, to highlight the essentially social nature of language. To use language is to be part of a group of language users (there can be no private language), so that the meanings of words and concepts found within these shared language-games are thus perspicuous and sound. It follows from this that what it means *to be creative* is in fact no different to what it *means* to be anything else in ordinary language. The term 'creative' has meaning through its use, and as a consequence it encompasses *all* of those things which I listed at the beginning of this paper, in varying degrees.

Since language is embedded in a way of life, and because what a word means is dependent on the language-game from—and within—which it derives meaning, it stands to reason that the component parts may change, and the balance will shift in favour of one aspect or another at any given time. Simply, creativity means different things in different language-games, and these games are linked by a notion of family resemblance (whereby two or more things can be connected by varying amounts of similarities). On this reading, we can easily accept that the type of creativity apparent in the composer example noted above is indeed *different* from that of the guided poet example, by virtue of many determinate and random factors. This is not to say however, that one should be pitted against the other with claims that one or other offers a more or less definitive account of creativity, but rather that they both share and diverge in different ways from a general notion of what it is to be creative. Any criteria we might give depends on the given individual language-game, and thus is open to change and regulation, in so far as normal language *always is*.

From this we might conclude that it is not an answer to the question posed in the title of this paper that we should seek, since the answer is likely to be negative, but rather a re-evaluation of what we hope to achieve in asking this question. Simply, it is not about asking whether a machine *is* creative, for the arguments I have offered make such claims impossible. A machine could not be creative *in our terms*, for the same reason that if a lion could talk, it could not speak our language because it does not occupy our form of life. Once this is accepted, and the temptation toward replication of human creativity is resisted, focus can instead turn to consideration of what it would mean for an AI program to be creative in *AI terms*. This need not mean any more than it does in the example given earlier of the composer and the poet. The composer might not be creative in the same way as the poet might claim to be, and yet they may still share some aspect of what it is to be creative in broader terms.

On this account, it seems that the question posed in this paper is not merely 'a philosopher's question' (Schank, 1988: 220), but is in fact a rather important one, and one which is open to all language users who seek to understand what particular words mean in particular situations. Accordingly, this question has as much significance for developers of AI as it does for philosophers. The replication of concepts will always require deeper analysis than the replication of objects, if only because there is no physical manifestation which can be consulted.

In this paper I have claimed that definitions are often too rigid, and are neither conducive nor helpful to understanding or making use of creativity. A more holistic approach to

understanding creativity as a family concept which gains meaning within particular language-games is offered as an alternative to this way of thinking. In a now famous example, Wittgenstein remarks that if we boil a person down to ash, this would not comprise all that that person is, or was (1966: 24). The point made illustrates my argument nicely, for even if we locate aspects of the creative within brain processes, or particular actions, this is not to define what creativity *is* in the same way that neurophysiological analysis that explains which parts of the brain govern language use will not tell us what language *is*.

Even if we replicate apparently creative processes in AI programs, and *even if* we create something which to all intents and purposes *appears* to have the same creative outputs, this in no way accounts for all aspects of what it means to be creative. Indeed, some aspects such as the aesthetic might prove to be elusive for no other reason than that aesthetics seem not to be bound by clear rules or boundaries. To be sure, there are certain methods that one might follow, guidelines for taste, and ways in which we can anticipate reactions. We might even be able to predict with some accuracy what will be successful (such as critics are often wont to do). Yet, this does not *fully* account for the apparently random nature by which we come to say of one thing that we like it, and of another that we don't, even where they might be very similar things. It may turn out that all such decisions are in fact as arbitrary as the mystery which we so frequently ascribe to them, but this would be beside the point.

The debate is a stimulating one and it seems clear that the nature of the creative is by no means settled. Yet, without this agreement it would seem impossible to ascribe to a being different to ourselves (by which is meant a non-human about which aspects of being cannot simply be taken for granted) the function of creativity. This need not, however, be a stumbling block for research into what creativity is, or even where it might be located. Nor should it provoke suspicion concerning the probability of success in the creation of any sort of artificial intelligence which replicates some aspect which we are willing to accept as AI creativity. Rather, what is key is to recognise that the application of the term 'creative' to a computer program, or even to ourselves, is to make use of a word which has meaning in very particular language-games.

On this account, *replication* (as a measure for success in AI development) is a limiting concept, and proves impossible for the simple reason that what creativity means is dependent on a (potentially unquantifiable) number of variables. That there are different forms of creativity *already* constitutes part of how we perceive creativity, and this argument might prove most fruitful for claims that aspects of non-human creativity—though they may be particularly or even substantially different from our own—should nevertheless be considered *creative* in some way. As AI programming becomes more sophisticated, and the data more extensive, it follows that the creative scope of an AI machine might move closer to what we understand our own creative abilities to be, all of which could be achieved without making metaphysical leaps. It is clear however, that even in spite of this, it is impossible that machine creativity (whether superior or inferior to our own) could ever be on a par with human creativity. My suggestion here is that coming to a better understanding of how creativity comes to have meaning will free AI developers from the need to try and *replicate* human creativity, if for no other reason that that would be as likely to succeed in this as they might in teaching a lion to talk.

REFERENCES

[1] Boden, M.A. (1999) 'Computer Models of Creativity' in Sternberg (1999), pp.351-372

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

[3] Boden, M.A. (2nd ed., 2004) *The Creative Mind: Myths and Mechanisms*, London: Routledge

[4] Bogousslavsky, J. (2005) 'Artistic Creativity, Style and Brain Disorders' in *European Neurology*, 54: 2, pp.103-111

[5] Bohm, D. and F.D. Peat (1987) *Science, Order, and Creativity*, London: Routledge

[6] Candy, L and E. Edmonds (1999) 'Introducing creativity to cognition' in *Proceedings of the 3rd Conference on Creativity & Cognition 1999*, pp. 3-6

[7] Churchland (2nd ed., 1988) *Matter and Consciousness: Contemporary Introduction to the Philosophy of Mind*, Cambridge, MA: MIT Press

[8] Dennett, D.C. (1995) *Darwin's Dangerous Idea: Evolution and the Meanings of Life*, London: Penguin

[9] Dutton, D. and M. Krausz (1981) *The Concept of Creativity in Science and Art*, The Hague: Martinus Nijhoff Publishers

[10] Edelstyn, N. M. J. (2007) 'Art, Creativity and Brain Damage in Artists' in *Cortex*, 43: 2, pp.282-4

[11] Feist, G.J. 'The Influence of Personality on Artistic and Scientific Creativity' in R.J. Sternberg (1999), pp.273-296

[12] Gabora Liane (2002) 'Cognitive mechanisms underlying the creative process', in *Proceedings of the 4th Conference on Creativity & cognition 2002*, pp.126-133

[13] Gardner, H. (1982) *Art, Mind and Brain: A Cognitive Approach to Creativity*, New York: Basic Books, Inc

[14] Heilman, K. M. (2005) *Creativity and the Brain*, NY: Psychology Press

[15] Johnson-Laird, P.N. 'Freedom and Constraint in Creativity', in R.J. Sternberg (1988), pp 202-219.

[16] King, M. 'The new metaphysics and the deep structure of creativity and cognition' in *Proceedings of the 3rd Conference on Creativity & cognition 1999*, pp.93-100

[17] Langley, P. and R. J. Jones (1988) 'A Computational Model of Scientific Insight', in R.J. Sternberg (1988), pp.177-201

[18] Lawson, B. '"Fake" and "Real" Creativity using Computer Aided Design: Some Lessons from Herman Hertzberger' in *Proceedings of the 3rd Conference on Creativity & cognition 1999*, pp.174-9

[19] Lawson, B. '"Fake" and "Real" Creativity using Computer Aided Design: Some Lessons from Herman Hertzberger' in *Proceedings of the 3rd Conference on Creativity & cognition 1999*, pp.174-9

[20] Martindale, C. (1999) 'Biological Bases of Creativity' in Sternberg (1999), pp.137-152

[21] Miller, D. L. (1989) *Philosophy of Creativity*, NY: Peter Lang

[22] Novitz, D. (1999) 'Creativity and Constraint', *Australasian Journal of Philosophy*, 77: 1, pp. 67-82

[23]Rothenberg, A. and C.R. Hausman (eds.) (1976) *The Creativity Question*, Durham NC: Duke University Press

[24] Saunders, R. and J.S. Gero, 'How to Study Artificial Creativity', in *Proceedings of the 4th Conference on Creativity & cognition 2002*, pp.80-7

[25] Schank, R.C. (1988) 'Creativity as a Mechanical Process', in R.J. Sternberg (1988), pp.220-238

[26] Sternberg, R.J. (ed.) (1999) *Handbook of Creativity*, Cambridge: CUP

[27] Sternberg, R.J. (ed.) (1988) *The nature of Creativity: Contemporary Psychological Perspectives*, Cambridge: Cambridge University Press

[28] Ward, T. B., S. M. Smith and R. A. Finke 'Creative Cognition' in. R.J. Sternberg (1999), pp.189-212

[29] Wittgenstein, L. (1966) *Lectures and Conversations on Aesthetics, Psychology and Religious Belief*, Oxford: Blackwell

[30] Wittgenstein, L. (3rd ed., 2001) (trans. G.E.M. Anscombe) *Philosophical Investigations*, Oxford: Blackwell

[31] Wittgenstein, L. (3rd ed., 1978) (trans. G.E.M. Anscombe) *Remarks on the Foundations of Mathematics*, Oxford: Blackwell

A Modelling Framework for Functional Imagination

Hugo Gravato Marques [1] and **Owen Holland** [2] and **Richard Newcombe** [3]

Abstract.

Imagination is generally regarded as a very powerful and advanced cognitive ability. In this paper we propose a modelling framework for what we call functional imagination: the ability of an embodied agent to simulate its own behaviors, predict their sensory-based consequences, and extract behavioural benefit from doing so. We identify five key components of architectures for functional imagination, and claim that they may be both necessary and sufficient. We outline a typical architecture, explain the flow of control within it, and describe a typical testing scenario using nested physics-based robot models. We also show how malfunctions within such an architecture may produce effects reminiscent of those found in certain human pathologies.

1 INTRODUCTION

Imagination has been regarded in Western philosophy as a very useful and significant cognitive ability. Prominent thinkers like Aristotle, Descartes, Hume, Kant and Sartre have all made contributions to the subject. In spite of this, however, little light has been shed on the mechanisms underpinning imagination [37]. One of the reasons why this might be so is that very often discussions about imagination are in fact discussions about imagery. It is true that the concepts cannot be completely dissociated, but it is important to be clear about the differences between them.

Imagery usually refers to an internally pictured object or situation ([38]); historically, it may be mentioned either in a phenomenological or a representational context. Phenomenologically, imagery is referred to as being capable of triggering experiences (or sensations) that resemble our experiences of daily life. This phenomenon is often called *quasi-perceptual* experience. The imagery of a dog can trigger a range of sensations one might have when one is actually close to a dog: the feeling of touching its rough coat, the sound of barking, or the dog-like smell. Representationally, imagery is strongly associated with an image-like representation ([38]). Imagination, on the other hand, is the process responsible for producing imagery. All references to imagination are references to things that are not present to the senses. However, some experiences of non-existent reality are not produced by imagination (e.g. the experience of an after-image or *phosphenes*). In addition, Penfield has shown that coherent complex experiences can be generated artificially simply by inducing currents in neurons located in specific areas of the brain ([26]).

Several theories have tried to explain the process of imagination through the phenomenological and representational aspects of imagery. The theories of Descartes and Hume are examples of such attempts. Descarte's dualist theory relied on a mystical soul to explain the process of imagination. Hume eliminated the reference to an immaterial entity by transforming the workings of the mind into a mechanical system ruled by laws of association, where each idea (in Hume an idea is basically an image) is brought to consciousness through the principles of resemblance, causality and contiguity.

We have discussed elsewhere several reasons why these two theories failed to explain the workings of human imagination ([16]). Here, it will suffice to say that it would be very hard to model the process of imagination solely in terms of the phenomenological or representational aspects of imagery simply because we are not yet at a stage of being able to access them scientifically. For this reason, we propose to focus on a third, and more recent, aspect of imagery - the neuroscientific aspect. We will try to identify what happens in the body and in the brain while imagery and imagination are occurring.

In neuroscience, several experiments suggest that roughly the same areas of the brain are active when a subject is sensing (or acting) overtly and when he doing so covertly. Although it is far from being universally accepted by neuroscientists, several experiments suggest the involvement of the early visual cortex during visual imagery [13]. In one experiment, subjects were asked to close their eyes and to visualise a set of previously seen striped patterns [12]. Results show that areas as early as area 17 and 18 of the visual cortex were active during imaging. In the same set of experiments the normal functioning of area 17 was disturbed in order to investigate the role it plays in imaging. Results showed that, by disrupting the normal operation of area 17, performance on the imagery (and perceptual) task was impaired.

Furthermore, recent studies of motor imagery suggest that motor imagery is functionally and anatomically related to motor execution. An fMRI study of finger movements showed significant activation of the SMA (supplementary motor cortex) and the PMC (premotor cortex) during both execution and imagery [14]. In the same study, the M1 (primary motor cortex) and S1 (somatosensory cortex) showed less activation during imagined finger movements. These results were confirmed by a similar study carried by Porro [27].

Psychological studies have also shown a striking relation between overt and covert behaviour. Shepard and colleagues

[1] University of Essex, UK, email: hgmarq@essex.ac.uk
[2] University of Essex, UK, email: owen@essex.ac.uk
[3] University of Essex, UK, email: ranewc@essex.ac.uk

have shown that the time taken to manipulate objects mentally seems to be linearly dependent on the number and extent of movements (or operations) made [30]. Subjects were asked to see whether, by folding a piece of paper in various indicated ways, two of its edges could be brought together. The results indicated that the time to reach a conclusion is dependent on the complexity of the folding process. The same sort of conclusion was reached in a mental rotation experiment where subjects were asked whether a given 3D object can be rotated to match another object [31]. The results indicated that the time to mentally rotate the image was proportional to the angle of rotation.

In addition, there also seems to be a connection between muscular activity during imagery and overt perception. In a study aiming at comparing eye movements during overt viewing and visual imagination, subjects were asked to look at the same irregularly-checked diagram four times (the diagram was rotated by 90 degrees between trials) while their eye movements were being recorded [2]. The eye movements were also recorded while subjects were imagining each of the four patterns. The results showed a close relationship between eye movements during overt and covert viewing suggesting also that eye movements reflect the content of what is being seen and what is being imagined. Similar results have been observed in different sensory modalities. An experiment testing the capabilities of subjects to imagine a smell have shown that it is very difficult (if not impossible) to image a smell without overtly sniffing [1].

All these results, and many more, suggest a strong relationship between imagination and the body, which highlights the possible relevance of embodiment theories to help explain this aspect of human cognition [40], [3]. However, it is only recently that ideas of embodiment have penetrated into Artificial Intelligence (AI). Previous attempts to capture the apparently abstract nature of human thought in AI were implemented within a symbolic framework. The General Problem Solver is an early and clear example of such an attempt (see for example [24], [25], [23]. In the General Problem Solver a group of operators could manipulate a collection of logical or mathematical expressions in order to find a solution to a given problem (e.g. a mathematical proof, or finding an analogy). The basic idea behind these reasoning systems relied on evidence from introspection and verbal protocols that humans seem to be able to manipulate propositions in order to form plans of action that can then be executed overtly. In spite of the failure of the Physical Symbol System Hypothesis to find support in neuroscience, logic-based approaches continue to dominate AI, but the lack of success in dealing with real world systems is finally turning attention towards the ideas of embodiment that are close to achieving dominance in cognitive science.

In this research programme, we want to use experimental results from neuroscience and psychology in order to produce a qualitative model (or candidate architecture) of human imagination. We do not aim to capture every single cognitive ability in which imagination might be involved, nor to account all the phenomena that seem to be related to imagination, such as dreaming, free-associative thought, etc. For this reason we will focus on what we call functional imagination: the mechanism that allows an embodied agent to simulate its own behaviours, predict their sensory-based consequences, and extract behavioural benefit from doing so [17] (see below). Through the construction of increasingly architectures from simple components we hope to be able (1) to show that they work, (2) to establish parallels between the way they work and experimental evidence from humans (or other animals), and (3) to investigate malfunctions in the models caused by missing or defective components and compare them, if possible, with disorders found in humans.

The remainder of the paper is structured as follows: in Section2 we will summarise some work in AI and robotics relevant to the topic of imagination; in Section3 we will outline the concept of functional imagination; in Section4 we will present one of our models for functional imagination; in Section5 we will describe an experiment using a complex and dynamic physics-based simulator to study the model; in Section6 we will make some qualitative comparisons between our model and some of the experimental data shown above; finally, in the Section7 we will make some concluding remarks.

2 CURRENT RESEARCH

Some research in AI is clearly relevant for a theory of imagination. Shanahan, for example, is creating architectures based on Baar's Global Workspace Theory that allow inner rehearsal of actions by a (computationally) embodied agent [29]. From the learned associations between sensorimotor patterns of neural activation, a simulated agent is able to extract the consequences of certain actions through inner rehearsal and select actions that lead to rewarding behaviour.

Ziemke and colleagues have been exploring the sharing of sensorimotor structures for driving a simulated Khepera robot around a room in the absence of sensory information. They implemented a wall following algorithm for driving a Khepera robot around a room in order to discretize the environment into different categories (e.g. corners, corridors, etc). At the same time they trained a recurrent neural network to predict the next category given the current one. Then by feeding the neural network with its own predictions they showed that the robot was able to 'imagine' itself driving around the room [33].

Using a broadly similar approach, Stein introduced MetaToto, an upgraded version of Mataric's robot Toto [19], which was able to navigate in an unknown environment and add new locations (nodes) to a dynamic map (graph). MetaToto was capable of goal-driven navigation using known landmarks; more interestingly for our concerns, the robot was able to move to new and unknown locations using very crude descriptions of the environment [32]. By reusing the mechanisms for sensing and acting, MetaToto was able to generate sensory representations of what it would be like to be in those places, and was thus able to find its way to unknown locations.

Mel's Murphy, a real robot equipped with an arm and a video camera, was able to solve grasping problems using visual imagery [22]. The robot worked in two modes. In the first mode it moved its arm around until it found a way to grasp an object. During this training stage, associations were created between the movements of the arm and the image of the arm and the object recorded from the camera. After the connections were established Murphy was able to 'imagine' the grasping of objects using only its (visual) imagery capabilities.

Embodiment is central to all these research projects. From a disembodied perspective Thaler claims to have invented a 'Creative Machine' that can perform discovery and invention at the human level in fields as diverse as drug invention, car design, dance steps, musical compositions, etc. [36]. The Creativity Machine has two main components: an Imagination Engine for generating new ideas and an Alert Associative Center, for evaluating the ideas coming from the Imagination Engine. The Imagination Engine is an Artificial Neural Network (ANN), which can be trained on some body of knowledge and then perturbed internally with just the right amount of noise. Creative ideas are supposed to come from the associations made during training combined with the right amount of noise added to the response of the ANN. Unfortunately Thaler's claims are very difficult to establish or assess from published data. For example, we could not find the mathematical parameters of the network providing the 'right' perturbation needed for potential ideas to arise, or any details on the way the evaluator actually operates. Other problems are the lack of external references to support claims such as: '[the] architecture emulates the thalamo-cortical loop in the brain (e. g., the seat of intelligence and consciousness) rather than blind [search]' [10].

3 FUNCTIONAL IMAGINATION

As mentioned before, in our project we are focusing on building architectures that can exploit neuroscientific data for producing architectures for functional imagination. We define functional imagination in the context of artificial embodied agents as the mechanism that allows an agent to simulate its own behaviours, predict their sensory-based consequences and extract behavioural benefit from doing so (see [17]). By behavioural benefit we mean an increase in reward or utility achieved by using internal simulation. Here, we present what we claim to be five necessary and sufficient conditions for the presence of functional imagination in any embodied agent. We will briefly discuss each condition in order to introduce some of the components in our models. We have demonstrated sufficiency elsewhere using a working implementation of a minimal architecture where the 5 conditions were included [17]; in this paper we will concentrate on the actual operation of a simple architecture.

Condition 1: Sensorimotor-based prediction

An embodied agent should be able to predict the consequences of its actions in terms of sensory-based activations. This idea has been advanced by [7] [8] and [9] and offers a possible explanation of neural activations in the sensory and motor areas of the brain during covert behaviour. This condition implies the existence of sensory-motor mechanisms as well as a mechanism for predicting the sensory consequences of a motor action. In control theory such a mechanism is called a forward model. In general a forward model is a mechanism that predicts the next state of any system (the plant) given its current state and the current action. Forward models have been argued to be very basic mechanisms that evolved initially for anticipation in motor control [7]. However the new idea is that, if detached from the external sensory data, the forward model can then predict the consequences of an action and substitute for the incoming sensory signal with its predicted value. If in addition to this an agent is able to select an action and inhibit it from overt execution, then the agent would be endowed with a sort of virtual world which could be detached from the external world, and in which actions could be tried covertly and their consequences predicted without further external information [6]. Dennett argues that an animal endowed with such a virtual world - a "Popperian creature" - would have an evolutionary advantage over other creatures, because it would able to try various risky 'hypotheses' of action without putting itself in real danger [4].

Condition 2: Goals

An agent must be able to execute goal-related behaviour. By goal-related behaviour we mean simply the ability to generate motor commands as a response to an internal state that might be changed as a result of the execution of those commands. This could be as simple as searching for food in response to hunger. In this situation the internal state of the agent is hunger (or some representation of the need for food), its goal is to reduce or eliminate hunger, the target of its action is food, and its behaviour is foraging.

McFarland distinguished between goal-directed, goal-achieving and goal-seeking behaviour [20]. A goal-directed system is one where the behaviour is guided by reference to an explicit internal representation of the goal to be achieved; for example, an explicit representation of the required percentage of stomach filling to be achieved. A goal-achieving system is one that can recognize the goal once it is arrived at (or that can at least change its behaviour once it reaches the goal), but where the process of achieving the goal is determined solely by the environmental circumstances. For example, an animal would keep foraging and eating until the stomach was full, when the signal from the full stomach would cause a switch in or cessation of the behaviour. Finally a goal-seeking system is one that is designed to approach the goal without the goal being explicitly represented within the system. A good example of this is a scheduling system that allocates a certain time slot for some particular behaviour (say, eating); when that time slot ends, some other behaviour is triggered.

For functional imagination, the goal is not required to be explicitly represented in the way required by a goal-directed system. Nevertheless, the goal needs to be recognized once the agent has arrived at it. The reason for this is that the usefulness of internal simulation must be measured in relation to a goal. Without the presence of a goal (be it implicit or explicit) internal simulation loses its functional value because there is no way of establishing whether it arrived at a useful result or not; it then becomes something closer to day-dreaming or the free association of ideas.

Condition 3: Evaluation

It is very hard (if not impossible) for an agent to behave in every situation in a way that maximizes its chances of survival and/or reproduction. For example, through lack of appropriate cognitive abilities, a dog might fail to identify the usefulness of a stick for taking food out of an otherwise inaccessible cage. If an agent was always able to behave optimally

according to some desired measure there would be no need for functional imagination. It is the possibility that the agent might produce behaviours that fail to achieve its goals that makes the role of functional imagination relevant. An agent therefore must in some way be able to evaluate its current state (be it real or imagined), which implies at least the capacity to distinguish whether a goal is fulfilled or not. As explained above this is a minimum requirement of any agent capable of either goal-directed or goal-achieving behaviour. Evaluations might be binary - stating simply whether a goal was achieved or not - or might take a range of values indicating the degree of satisfaction of the goal according to some measure (say, energy expenditure).

Condition 4: Action selection

An agent must be able to select actions (or motor responses) for internal simulation. Animals have a number of different tasks that have to be performed in order to enable their survival and reproduction (e.g. eat, drink, mate, etc). This means that they must be capable of producing different and appropriate behavioural responses in order to fulfil each task. This action selection is also necessary for dealing with simulated actions. If an animal was able to perform only one possible action, the usefulness of imagination would be restricted to the decision of executing the action or not, according to the evaluation of the consequences of the simulated action. In all other cases, functional imagination requires the agent to be able to try different actions in the same situation. As a minimum, this demands that the same action should not be repeated even though the state remains the same; a simple mechanism implementing inhibition of return [11] can fulfil this requirement.

Condition 5: Selection of sensorimotor-based state

An agent must be able to imagine situations that are not tightly tied to its current context. Selecting the scenario within which internal simulation is performed is essential to allow the agent to set its internal state independently of its current state. This would also allow the agent for example to simulate what it would have happened in a past situation if some other actions had been taken - enabling reflection and enhanced planning [34]. In addition, different states are produced during internal simulation as a result of simulating different actions and it will be useful - for example, in multi-step planning - for the agent to be able to select the state within which actions will be simulated. Without this mechanism an agent would be restricted to scenarios based only on its current state.

4 ARCHITECTURE

In our project to date we have implemented a variety of architectures. The reason for this is simply that we suspect (and therefore we want to show) that increasing imaginative ability comes at the expense of an increasing number and variety of components in the architecture. In order to differentiate and categorize these architectures we have created a taxonomy. We differentiate between architectures that reuse the same sensory-motor structures for both overt and covert behaviour (economical architectures) and architectures that use copies of those structures for covert behaviour (duplicated architectures).We differentiate between architectures that overtly trigger the first solution they find for a certain problem (reactive architecture) from architectures that are capable of applying the best solution found within a given time (rational architecture). We also differentiate between architectures that can simulate only one step ahead (single-step architectures) and those that can cope with several steps ahead (multi-step architectures). Finally, we distinguish between architectures that retain and use memories of previous plans (memorizing architectures) and architectures that have to search afresh for a plan every time a given problem appears (memoryless architectures). Due to space limitations, we will present only one architecture here, an example which in our taxonomy would be classified as reactive, single-step, economical, and memoryless).

4.1 Architecture components

The architecture we will describe here is shown in Figure1. When producing a model of functional imagination, one of the first questions to be answered is how the system can distinguish reality from fiction (what is being imagined) [28]). In our architectures we use a switch mechanism both to implement and to capture this distinction. As can be seen in Figure1 the switch mechanism affects both the sensory and the motor systems. When the switch mechanism is set to ON the sensory system uses the information coming from the real world; when the switch is set to OFF the sensory system uses the information provided by the forward model. In addition, when the sensory system is set to ON the motor actions are executed overtly, while when it is set to OFF any motor action is inhibited from overt execution. In between the sensory and motor systems there are two selection mechanisms: one to select an action for execution, and another for selecting the feature (target) at which the action will be directed.

The short-term memory connected to sensory system (*State0 STM* mechanism) allows the agent to set the scenario (sensory state) that will be used as the starting point for the plan. In this implementation this state is the sensory state at the time the switch is set to OFF, but in other architectures this is not necessarily the case. In addition, because this is a single-step architecture, the state stored in this memory will be the starting point for the simulation of every action tried; it must replace the sensory states produced by the forward model after each simulated action.

The sensory mechanism is connected to an evaluation mechanism which allows the agent to evaluate its current state in relation to its current goal. In addition, the evaluation mechanism determines whether the current action and feature selection policies should be changed or not. A policy is a mapping from sensory states to the actions or plans that the agent ought to perform ([35]). If the evaluation is positive then the target and action policies are changed in order make the decisions that led to the rewarding state more salient and more likely to be chosen in the future. If, on the other hand, the evaluation is negative the policy should be changed in order to allow for other actions to be preferentially selected. In this architecture a plan entails only one action because it is a

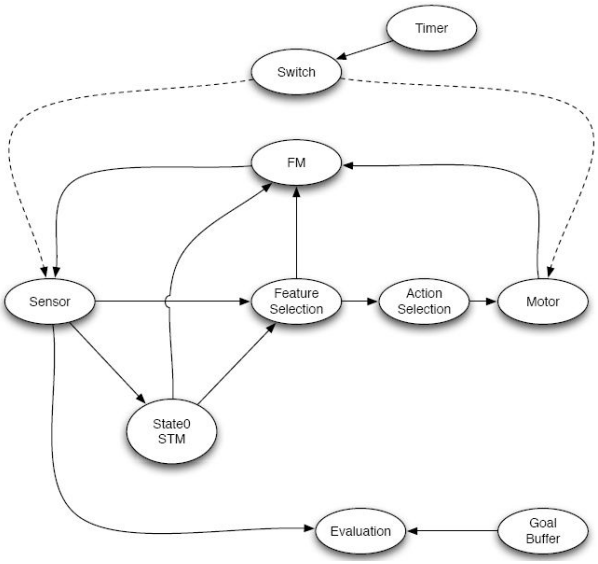

Figure 1. The architecture implemented.

single-step architecture.

Finally, we can see that the switch is connected to a timer. This allows the architecture not only to allocate the amount of time it will spend searching covertly for an action or plan, but also to interrupt the covert simulation (by setting the timer to OFF) in response to any unexpected event that might happen elsewhere in the system.

4.2 Control Flow

In addition to the components of the architecture we also present here the control flow of the activities in the system (see Figures 4 and 5). The architecture is divided into two main functional groups: one responsible mainly for controlling the switch (Figure4) and another responsible for controlling the behaviour (overt and covert) of the agent (Figure5). Each group works on its own internal loop but one is capable of influencing the other. A good analogy here is the way two threads in an operating system can run their own internal loops independently, but each influences the other by (for example) setting common variables.

In the implementation of the architecture the agent starts by being engaged in overt behaviour under the control of the sensory and motor policies until a goal is set in the system. In the current implementation the goal is set manually, but in an agent fully embodied in the world the goal could arise automatically - for example, the need for food, or for a mate. Once a goal arises in the system the agent performs the necessary steps for entering simulation mode: it stores the current state in the *State0 STM* mechanism, sets the timer to the amount of time allocated for solving that specific problem, and sets the switch mechanism to OFF, which inhibits the sensory and motor components from communicating with the external world. In this implementation the timer is set to a fixed time, but in an agent fully embodied in the world the time allocation for solving a problem should be dependent on some measure of significance of the goal in the current context of the agent. Time allocation is an essential ability of any embodied agent that must fulfil multiple goals in order to survive [21] and the allocation of time to simulate and imagine is just an extension of that idea. After the initialization is complete this functional group does not do anything until the time for solving the problem expires.

Once the switch is set to OFF the architecture responsible for the behaviour stops whatever action was running and evaluates the current sensory state. From this point on this sub-architecture simulates actions covertly until the switch is turned ON. The simulation of one action starts by loading the sensory state stored at the time the problem arose. As mentioned above, the reason for this is that the architecture presented here is a single-step architecture (see above). After loading the state, the agent selects a target and an action and simulates the result of executing that action using the forward model. Once the action is complete the agent evaluates the resulting state; if the state achieves the current goal then the agent sets the timer to OFF and sets a flag indicating that it found a solution for the goal; otherwise it simulates another action.

Once the timer is OFF (either because the agent found a solution for its problem or because the time for solving the problem expired) the architecture responsible for controlling the switch sets the switch to ON. This will force the sensing mechanisms to receive information from the real world. In an embodied system, if no solution was found during the simulation period the agent should choose what to do next based on its current context (e.g. take more time to simulate, or perhaps do nothing until the situation changes); here, the architecture simply terminates stating that no solution was found. If, however, a solution has been found, the policies for the target and action selection are reinforced to bias them in favour of executing the action overtly.

Once the overt execution of the plan (one action) terminates, the final state of the (real) world is evaluated. If the evaluation is positive the problem is solved; otherwise, the architecture states that the plan was unsuccessful. Here, as before, the current context of the agent and the nature of the problem should determine what happens next.

For reasons of clarity, there is a part of the diagram that is not included in Figure5, which is connected to the decision point "Action complete" (see Figure2). This part deals with the problem of what happens if a goal cannot be achieved because the action in the plan was for some reason unsuccessful (e.g. perhaps due to a sudden change in the world). If an action during the execution of a single step or multi step plan was unsuccessful the plan becomes obsolete and the architecture terminates in a state of 'plan unsuccessful'. Once again, it should be the current context of the agent that determines what to do next.

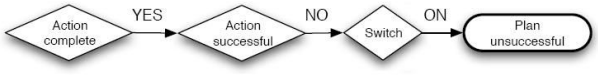

Figure 2. Part of the work flow that was not included in Figure 5.

5 EXPERIMENT

In order to test our architecture, we used our physics-based humanoid simulator - SIMNOS [5]. In SIMNOS's body the skeletal components are modelled as jointed rigid bodies, with spring-damper systems at each joint. The rigid body limbs are fully contactable surfaces which allow the robot to interact with its environment. Each muscle is modelled as a single parallel spring-damper system with asymmetrical conditioning of the spring and damper constants, in order to allow the muscle to produce force only when is contracted (for more details see [5] and [18]). The robot has a monocular visual system (first-person view) that allows it to capture coloured images of its environment in a way similar to a camera mounted on a real robot. This simulated camera also provides the agent with the distance of each projected pixel.

For this experiment we used two instances of SIMNOS running in parallel; one to capture the interactions between the *real* agent and the *real* world, and the other to capture the covert interactions of the agent (the result produced by the forward model in the internal architecture). We will call the former the *real agent* and the latter the *virtual agent*. A similar approach where a second instance of a simulator is used as an internal model of the first instance can be found in [39].

5.1 The task

In our experimental setup, the real agent has in front of it a blue object and a red object on a table top (see Figure3). As can be seen in Figure3, the same general scene is loaded into the virtual agent, but the objects are not included. The real agent can visually explore the real environment by moving its head around. Every time an unknown object is found in the real world the virtual environment is updated by placing an object of the same colour in the appropriate position. The objects are distinguishable by their colour.

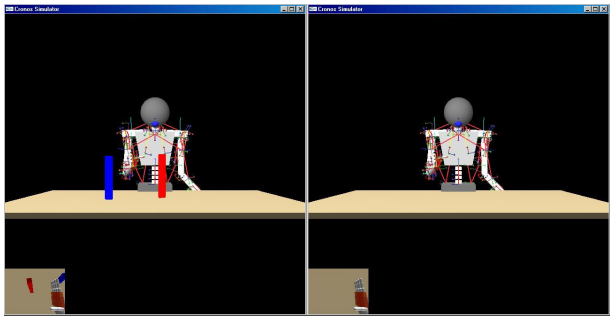

Figure 3. The experimental setup: the *real* world on the left and the virtual world on the right. At the beginning of the task the agent's virtual world (right) does not contain objects to interact with; they are added as the agent explores the real world.

Before we set the goal manually, the agent was given some time to explore its environment - enough to find the two objects and update its virtual world. The goal the agent was required to achieve was to move the red object further than a certain fixed distance. In order to solve the task the agent was endowed with two pre-programmed behaviours: one that allowed it to grasp an object, and another that allowed it to grasp an object and throw it forward. The goal distance threshold was set to a value that could usually be exceeded when the agent executed the throwing behaviour on any of the objects. Once the goal was set, the architecture ran as described above until one of the end states was reached.

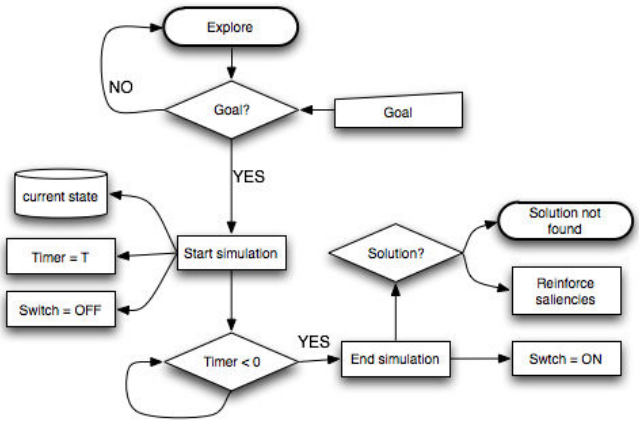

Figure 4. Control flow of the architecture responsible for controlling the switch.

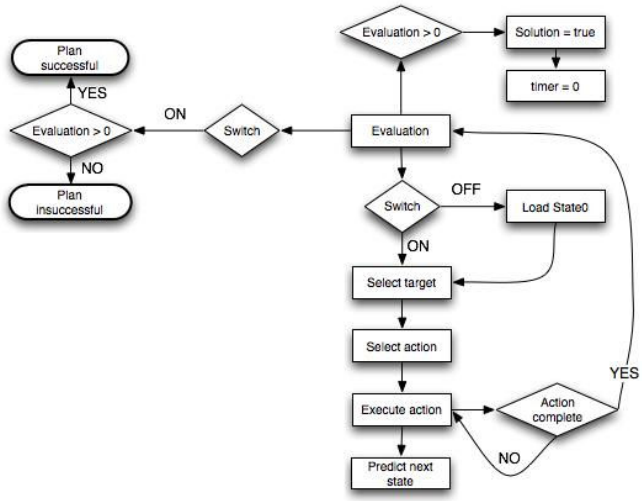

Figure 5. Control flow of the architecture responsible for the behaviour of the agent (overt and covert).

5.2 Selection mechanisms

We adopted a reinforcement learning strategy in order to form the target and action selection policies appropriate for reaching the goal. The reinforcement learning algorithm had to arrange that the red object and the throwing behaviour should be more salient that the blue object and the grasping behaviour respectively. This would allow the agent to throw the

red object forwards and exceed the distance threshold. In this single-step architecture the target and action saliences were modelled using two arrays, one for each selection mechanism. Because there were two objects, the feature salience array contained two values, one for each object. The same applied to the action salience array.

The salience arrays were initialised with low random values. Every time a negative evaluation was received (meaning that the goal had not been achieved) the feature and the action just selected were punished by decreasing the salience of each by a fixed amount. During the search, the probability of selecting the most salient item (feature or action) was set to a value of 0.75[4]. This meant that 25% of the times the item selected was the less salient one.

5.3 Results

The results of the experiment showed that the agent was reliably capable of solving the problem overtly when given enough time to search covertly. The behaviours implemented are not certain to be successful, and occasionally fail. For example, the agent might fail to grasp an object, or might drop the object when trying to throw it. Uncertainty is a feature of complex and dynamic environments (such as the real world, and also our simulator) and this is actually a good strategy for testing our model; it shows that it is capable of coping with deviations from perfection. In addition, the architecture was also able to detect when a solution was not found within the time slot allocated.

6 DISCUSSION

6.1 Activation of sensorimotor areas during covert behaviour

We started this paper by mentioning some experimental results showing the relations between overt and covert behaviour at the level of brain and behaviour. One of the themes was the appropriate activation of sensory and motor areas during imagery. In our model we have shown that executing actions covertly necessarily activates the sensorimotor mechanisms active during overt behaviour.

6.2 Overt and covert behaviour

Other results have shown that the time for performing a mental rotation depends linearly on the angle of rotation, and that the time taken to imagine folding a piece of paper increases with the number of operations that need to be performed. In our architecture, the time that it takes to imagine grasping a target depends on the distance of the target. Here, one could argue that by using a second instance of the same simulator as the forward model the time to reach a target overtly and covertly should actually be the same. This is true if one assumes the simulators run at the same speeds, which in our implementation they do. However, the linear dependence between the distance to the target and the time to grasp it would happen in any sort of model that could covertly produce data that is qualitatively similar to the real data. At the moment we are at a stage where we will start implementing the architectures in our real humanoid robot. Once this is done we will have more data to prove our point. For the same reason, time that it would take to imagine the execution of several behaviours would be dependent on the number of behaviours that need to be executed. Here, we only present data regarding a single behaviour, but this can be shown in the implementations of our multi-step architectures.

[4] This value is almost irrelevant in the current implementation as there are only two choices per selection mechanism. The only situation we want to avoid is that of a cycle where *feature1* and *behaviour1* are selected first, then *feature2* and *behaviour2*, and then back again to *feature1* and *behaviour1*. The use of a stochastic method for selecting the most salient behaviour ensures the diversity of the behaviour.

6.3 Defects and pathologies

The simple architecture we have described has a key component - the switch. It can malfunction in a number of ways, and some of these have intriguing parallels to some human pathologies. For example, it is supposed to be OFF during episodes of internal simulation; if it fails to operate, the imagined behaviour will be executed in reality. In humans, motor commands produced during dreaming are normally inhibited, but in some people, popularly called sleepwalkers, actions corresponding to dream elements are occasionally carried out, and several cases of murder have been defended in the American courts by claiming that the alleged murderer had been unconsciously acting out a dream. One eventual aim of this project is to systematically damage the architectures produced and to log the system pathologies, with the aim of comparing these to standard databases of human pathologies. Correspondences will not amount to proof that our models capture the processes of human imagination, but the comparison will at least reveal something about the goodness of fit.

6.4 Simultaneous overt and covert behaviour

We have argued elsewhere [15] that an architecture that reuses its sensory and motor systems for overt and covert behaviour cannot act covertly at the same time as it acts overtly . In fact, the question of whether we reuse the exact same neural structures for overt and covert behaviour or whether we use copies of those systems (which must be localized in the vicinity of the sensory and motor areas for the experimental results to hold) is an open question. We have already implemented architectures that deal with this problem, and allow an agent to act covertly and overtly at the same time and we hope to publish the results in the near future.

7 CONCLUSION

In this paper we have proposed a modelling framework for what we call functional imagination: the ability of an embodied agent to simulate its own behaviors, predict their sensory-based consequences, and extract behavioural benefit from doing so. We have identified five key components of architectures for functional imagination, and claim that they may be both necessary and sufficient. We have outlined a simple architecture, explained the flow of control within it, and described a

typical testing scenario using nested physics-based robot models. We have also speculated about how malfunctions within such an architecture may produce effects reminiscent of those found in certain human pathologies. This is ongoing work, as yet in its early stages, but it holds some promise of leading to a systematic synthetic approach to understanding the problem of imagination.

8 ACKNOWLEDGEMENTS

We would like to thank the Portuguese FCT for the PhD fellowship to Hugo Gravato Marques and to the EPSRC (GR/S47946/01) for funding the CRONOS project.

REFERENCES

[1] M. Bensafi, J. Porter, S. Pouliot, J. Mainland, B. Johnson, C. Zelano, N. Young, E. Bremner, D. Aframian, R. Kahn, and N. Sobel, 'Olfactomotor activity during imagery mimics that during perception', *Nature Neuroscience*, **6**, 1142 1144, (2003).

[2] Stephan Brandt and Lawrence Stark, 'Spontaneous eye movements during visual imagery reflect the content of the visual scene', *Journal of Cognitive Neuroscience*, **9**, 27–38, (1997).

[3] Ronald Chrisley and Tom Ziemke, 'Embodiment', in *Encyclopedia of Cognitive Science*, 1102–1108, Macmillan Publishers, (2002).

[4] Daniel Dennett, *Darwin's Dangerous Idea*, Harmondsworth, Allen Lane The Penguin Press, 1995.

[5] David Gamez, Richard Newcombe, Owen Holland, and Rob. Knight, 'Two simulation tools for biologically inspired virtual robotics', in *Proceedings of the IEEE 5th Chapter Conference on Advances in Cybernetic System*, pp. 85–90, Sheffield, (2006).

[6] Rick Grush, 'An introduction to the main principles of emulation: motor control, imagery, and perception', Technical report, UC San Diego, (2002).

[7] Rick Grush, 'The emulation theory of representation - motor control, imagery, and perception', *Behavioral and Brain Sciences*, **27**, 377–442, (2004).

[8] Germund Hesslow, 'Conscious thought as simulation of behaviour and perception', *Trends in Cognitive Science*, **6**(6), (2002).

[9] Owen Holland and Rod Goodman, 'Robots with internal models: A route to machine consciousness?', in *Machine Consciousness*, ed., Owen Holland, Imprint Academic, Exeter, UK, (2003).

[10] Imagination Engines Incorporated. Iei's patented creativity machine, 2005. [Online; accessed 27-August-2007].

[11] Laurent Itti and Christof Koch, 'A saliency-based search mechanism for overt and covert shifts of visual attention', *Vision Research*, **40**, 1489–1506, (2000).

[12] Stephen Kosslyn, 'The role of area 17 in visual imagery: Convergent evidence from pet and rtms', *Science*, **284**, 167–170, (1999).

[13] Stephen Kosslyn, Giorgio Ganis, and William Thompson, 'Neural foundations of imagery', *Nature Reviews Neuroscience*, **2**, 635–642, (2001).

[14] Martin Lotze, Pedro Montoya, Michael Erb, Ernst Hulsmann, Herta Flor, Uwe Klose, Niels Birbaumer, and Wolfgang Grodd, 'Activation of cortical and cerebellar motor areas during executed and imagined hand movements: An fmri study', *Journal of Cognitive Neuroscience*, **11**(5), 491–501, (1999).

[15] Hugo Marques and Owen Holland, 'Minimal architectures for embodied imagination', in *In Proceedings Brain Inspired Cognitive Systems (BICS2006)*, (2006).

[16] Hugo Marques and Owen Holland, 'Architectures for imagination: why models matter', (2007). Unpublished.

[17] Hugo Marques and Owen Holland, 'Architectures for embodied imagination', *Neurocomputing*, (2008). Submitted.

[18] Hugo Marques, Richard Newcombe, and Owen Holland, 'Controlling and anthropomimetic robot: A preliminary investigation', in *Proceedings of ECAL2007.*, Lisbon, (2007). Springer Verlag.

[19] Maja Mataric, 'Integration of representation into goal-driven behaviour based robots', *IEEE Transactions on Robotics and Automation*, **8**(3), 304–312, (1992).

[20] David McFarland, 'Goals, no-goals and own goals', in *Goals, No-Goals and Own Goals: A Debate on Goal-Directed and Intentional Behaviour*, eds., Alan Montefiore and Denis Noble, Unwin Hyman Ltd, London, (1989).

[21] David McFarland and Thomas Bsser, *Intelligent Behavior in Animals and Robots*, The MIT Press, 1993.

[22] Barlett Mel, 'Murphy: A robot that learns by doing', in *Neural information processing systems*, American Institute of Physics, New York, (1988).

[23] Allen Newell and Simon Herbert, 'Gps, a program that simulates human thought', in *Computers and Thought*, eds., Edward Feigenbaum and Julian Feldman, McGraw-Hill, (1963).

[24] Allen Newell, J. Shaw, and Herbert Simon, 'Report on a general problem-solving program.', Paris, (1959). Proceedings of the International Conference on Information Processing.

[25] Allen Newell, J. Shaw, and Herbert Simon, 'A variety of intelligent learning in a general problem solver', in *Self Organizing Systems*, eds., Yovits and Cameron, Pergamon Press, (1960).

[26] Wilder Penfield, 'Some mechanisms of consciousness discovered during electrical stimulation of the brain', *Proceedings of the National Academy of Sciences*, **44**(2), 51–66, (February 15 1958).

[27] Carlo Porro, Maria Francescato, Valentina Cettolo, Mathew Diamond, Patrizia Baraldi, Chiava Zuiani, Massimo Bazzocchi, and Pietro Prampero, 'Primary motor and sensory cortex activation during motor performance and motor imagery: A functional resonance imaging study', *The Journal of Neuroscience*, **16**(23), 7688–7698, (1996).

[28] Jean-Paul Sartre, *Imagination: A Psychologic Critique*, The University of Michigan Press, 1962.

[29] Murray Shanahan, 'Cognition, action selection, and inner rehearsal', *Proceedings IJCAI 2005 Workshop on Modelling Natural Action Selection*, 92–99, (2005).

[30] Roger Shepard and Christine Feng, 'A chronometric study of mental paper folding', *Cognitive Psychology*, **3**, 228243, (1972).

[31] Roger Shepard and Jacqueline Metzler, 'Mental rotation of three-dimentional objects', *Science*, **171**, 701–703, (1971).

[32] Lynn Stein, 'Imagination and situated cognition', Technical Report 1277, MIT AI Lab, (1991).

[33] John Stening, Henrik Jacobson, and Tom Ziemke, 'Imagination and abstraction of sensorimotor flow: Towards a robot model', in *Proceedings AISB2005 Symposium on Next Generation Approaches to Machine Consciousness*, pp. 50–58, Hatfield, UK, (2005).

[34] Susan Stuart, 'The binding problem: Induction, integration and imagination', in *AISB2005: Proceedings of the Symposium on Next Generation Approaches to Machine Consciousness*, Hatfield, UK, (2005).

[35] Richard Sutton and Andrew Barto, *Reinforcement Learning*, The MIT Press, 1998.

[36] Stephen Thaler, 'Neural networks that autonomously create and discover', *PC AI*, (1996).

[37] Nigel Thomas, 'Imagining minds', *Journal of Consciousness Studies*, **10**(11), 79–84, (2003).

[38] Nigel Thomas, 'Mental imagery', in *The Stanford Encyclopedia of Philosophy*, ed., Edward N. Zalta, (2005).

[39] Richard T. Vaughan and Mauricio Zuluaga, 'Use your illusion: Sensorimotor self-simulation allows complex agents to plan with incomplete self-knowledge', (September 2006).

[40] Tom Ziemke, 'What's that thing called embodiment?', in *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, Mahwah, NJ, (2003). Lawrence Erlbaum.

The Plaited Structure of Time in Information Technology

Ganascia Jean-Gabriel
LIP6, University Pierre and Marie Curie
104, avenue du Président Kennedy, Paris, France
email: Jean-Gabriel.Ganascia@lip6.fr

Abstract. The aim of this paper is to try to understand the structure of time in information technologies. Starting with historical arguments, it first shows that time is neither linear nor cyclic. This becomes clear if we consider that many technologies which were developed in the past and then gradually ignored have since enjoyed revivals. In a way, information technologies seem to have a life cycle. On the other hand, the time of information sciences and technologies is anything but cyclic, because information sciences and technologies introduce radically new devices that change our lives. One of the consequences of this strange structure of time is that, very often, long term predictions are more reliable than short term forecasts. This paper tries to explain the entangled structure of time in information sciences and technologies through a cybernetic model. It argues that the differences between information sciences and technologies and other technologies are due to their interactivity, whereby society intervenes in the design process.

1 Introduction

It is now a commonplace to talk about the acceleration of time and increasing world complexity. Everyone agrees that science and technology are producing ever more new results at an ever faster pace. It would thus seem to be becoming more and more difficult to make reliable predictions over a long period. However, in the field of IST — Information Science and Technology — things are not so simple and, surprising though this may be, the near future is more difficult to predict than the long-term future. We just need to look at the history of IST, which shows that short and medium-term predictions of, say, five to seven years have very often turned out to be wrong, while many long term forecasts have come true. For instance, let us take the example of the history of cybernetics, machine translation, artificial intelligence, expert systems, virtual realities, communication networks, minicomputers, microcomputers, the web, etc. We have regularly been wrong about the evolution and social consequences of these technologies. Meanwhile, many long-term predictions, 20, 30 or 40-year forecasts, say, have been correct. To be convinced, it is sufficient to read Alan Turing's forecasts in 1950 (cf. section 3.2), in his famous paper [12] about machine intelligence, or Jean-Francois Lyotard's 1979 book [2] relative to the status of knowledge in post-modern societies (cf. section 3.4) etc. The difficulty when trying to forecast imminent change comes simultaneously from the large number of different people taking part in the decision-making process and from the generalization of interaction that is typical of modern societies. Social interaction is becoming crucial in the development of advanced technologies — especially in the field of IST — and social practices play a more and more important role in the design process. The success of "user-centred design" (cf. [7]) in information technology is a sign of this trend. Since feedback is continuous and modifies predictions and expectations, production planning is becoming far more difficult and short term predictions are very risky. The structure of time is thus changing today and, as we shall see, is neither linear, nor cyclic, as was the case in the past. The aim of this paper is to try to understand this structure. To do so, it compares examples taken from the history of IST development with earlier predictions, and then suggests a cybernetic model in order to understand these changes to the time structure. In addition to this introduction, the paper is divided into four parts. The first shows how wrong past short-term predictions in the field of IST were; the second looks at medium and long-term predictions; the third provides an explanation of this phenomenon, and the fourth and last one proposes some conclusions about the new structure of time, which results from the broad dissemination of information and communication technologies.

2 Short term predictions

The history of data processing is marked by a long succession of errors of appreciation. The majority of short or medium-term forecasts were completely erroneous. To be convinced of this, it is enough to look at the major stepping stones in the development of data processing and, in particular, the software technologies. There are countless examples, of which a few are given below.

2.1 Cybernetics

At the beginning of the forties, cybernetics was enthusiastically embraced by many researchers who thought this offered radically new perspectives. The first automaton networks imagined by Warren McCulloch and Walter Pitts (cf. [3]), as well as the notions of feedback and of teleological machines introduced by Arturo Rosenblueth, Norbert Wienner and Julian Bigelow (cf. [10]), seemed to open up new horizons. According to many, a bridge between engineering sciences on the one hand and the brain or social simulation by means of information flows on the other, was in the process of being born. Many people thought they were now going to be able to establish the laws of complexity (cf. [6]) that govern biological phenomena, in particular the functioning of living organisms; physical phenomena, for example the spontaneous organization of atoms in crystals; political and social phenomena, which would make it possible to introduce

laws of government based on strict foundations and, finally, neuron phenomena, which produce thought. The first attempts were quickly followed by disappointment. For example, Frank Rosenblatt's attempts to characterize an artificial retina (which was called the PERCEPTRON (cf. [9])) encountered many pitfalls, which were underlined at the end of the sixties by Marvin Minsky and Seymour Papert (cf. [5]). Later, at the beginning of the eighties, many paths of research which had been opened up by cybernetics, and which had seemed to lead to dead ends, were reopened. What we have here is an example of technological revival, as mentioned in the introduction. However, as we shall show in the following, this is far from being the only case.

2.2 Machine Translation

At the beginning the fifties, considerable sums of money were invested in machine translation projects. Many people believed that computer resources, in particular storage capacity and their ability to handle character strings, would make it possible to build, quickly and cheaply, a machine able to automatically translate texts from one language to another, for instance texts written in Russian or French would automatically be transcribed in English. In support of this hypothesis, there was the possibility of designing dictionaries. Today, it seems obvious that it is not enough to have a dictionary in order to translate correctly; it is above all necessary to understand, which supposes a syntactical parsing followed by a semantic analysis and then a pragmatic analysis, many phases of which the engineers thought could be skipped. As a consequence, they were quickly disenchanted. However, these general considerations did not prevent the American government from engaging important sums of money in machine translation, and it also was the case of many European governments. It was only in 1966 that a report commanded by the American Senate [8] recognized the impasse at which the efforts of machine translation had arrived. As a result, the funding of all American laboratories working on machine translation quickly dried up and they were closed. However, this did not put an end to machine translation. In Europe, some laboratories continued to do research in this field and even today there are European research programmes, funded by the European Community, that support fundamental research on machine translation. Machine translation systems are even available to the general public on the Internet. The ambitions of machine translation have been reduced; it is no longer a question of translating all texts from one natural language to another, but of translating technical papers within a narrow field of knowledge, where there are no semantic ambiguities. The methods used now require knowledge of linguistics and semantics. Nevertheless, whatever the future of machine translation, the expectations and predictions of the early fifties now seem totally justified, even if they appeared to be completely wrong from the end of the sixties to the mid eighties.

2.3 Artificial Intelligence

From its very beginning, in 1956, artificial intelligence raised many hopes. It was imagined that a machine would be able to simulate almost all the cognitive capacities of intelligent beings (cf. [4][1]); for instance, that a computer would be able to perceive the outside world, to argue just like a human being, to play chess and to beat the best world players, to speak different languages and to understand everything. The very first research fed such expectations: a machine automatically proving most of the theorems of logic contained in Alfred Whitehead and Bertrand Russell's "Principia Mathematica" was built. These successes encouraged the pioneers of artificial intelligence to go further; they began to dream. So, Herbert Simon, future Nobel prizewinner for Economics and who also went on to receive the Turing awards, made resounding statements with his colleague Alan Newell (cf. [11]). According to them, 10 years hence (we were in 1958), if they were not excluded from international competitions, computers would no doubt become the world chess champions. Similarly, a computer would certainly be capable, again within 10 years, of composing music endowed with an unmistakable aesthetic value, of demonstrating totally original mathematical theorems, of imitating the psyche to the point that all psychological theories would have to be expressed by the means of computer programs, etc. It goes without saying that these predictions were soon proved wrong and, in 1965, a chess computer program was defeated by a 10-year-old child. Nevertheless, in 1997, that is to say less than 40 years on, a computer succeeded in challenging and beating the world champion chess player; computers are used a lot by musicians; computers play an important role in the activity of mathematicians to demonstrate new theorems; and psychologists also use many computer models. All of which goes to show that most of these announcements, even if they were proved wrong at the time, were not totally absurd. But the periods of time were not respected. Still in the field of artificial intelligence, many people got very excited at the beginning of eighties about what were called "expert systems". These are pieces of software which, unlike traditional computer programs, include specialized knowledge held by experts. Here, too, expectations were frustrated: the industrial development of expert systems or knowledge-based systems was much slower than what all specialists or forecasters had imagined. However, today, many industrial applications of these technologies can be found.

2.4 Computer Networks

More generally, for anyone prepared to examine it closely, the history of data processing contains a surprising succession of forecasting errors. In the sixties, experts imagined that data processing would develop on a centralized model, with a huge mainframe computer to which everybody would be connected. For many engineers of that time, computing capacity should be viewed as water or electricity flow, with a central production source to irrigate a whole company, even a whole city or a whole state. As a consequence, places such as office blocks or universities which were built at that time were cabled so that everyone would have direct access to the computer resources from their own offices. By the end of the seventies, minicomputers had been developed, which meant that data processing could be decentralized, i.e. each department of a big company would have its own computer. In this context, connecting all the offices to a central computer network was no longer justified. All that was required was a local area network so that each department could access the appropriate local computer resources. A few years later, workstations appeared on the scene: these were very expensive personal computers that experts thought engineers could be equipped with. These personal computers were closely followed by microcomputers, in other words desktop computers, whose relatively low cost changed the game completely. Not only did all the engineers and administrators have a computer on their desks, but so did all the secretaries. As a consequence, it was necessary to set up networks to connect all the microcomputers. Note that these changes were not at all predictable and it took IBM, the largest office computer company, quite some

[1] The original text can also be found at $http://www-formal.stanford.edu/jmc/history/$

time to be convinced of the place of microcomputers in offices. The first Personal Computer, i.e. the first IBM microcomputer, was only made and marketed in 1981, that is to say more than nine years after the first personal computers were designed and four years after the commercial success of Apple II. In this respect, the history of human-computer interfaces shows the hesitations before widespread use of the mouse and the adoption of the desk, files and trash can or waste bin metaphors. For a long time, the big computer firms considered that computing was only the business of specialists and no one even contemplated making computers accessible to the general public. Developed by Xerox Park as early as 1972, the first machines destined for non-specialists, in particular the Viola, the Dolphin and the Star, included a mouse and a high definition screen. However, they were not a commercial success, and it was the same thing with Lisa, the first machine developed in 1983 by Apple along the same lines. It was only in 1984, with the appearance of Macintosh, that these new human-computer interfaces conquered the world.

2.5 Internet and the Web

It is the same story for the telecommunication networks between computers. Remember that the first attempts at digital communication started at the end of the sixties. Development of the ARPANET network began at the beginning of the seventies in the United States, while at the same time similar projects were being developed in France, the Cyclades project, for example, or the Minitel, which relies on the use of the telephone network and which was operational at the beginning of the eighties. Many people spoke about on-line data processing. All the principles of networks were already there. And the coupling of computer networks and telephone networks, very much part of things today, already existed. Remember too, that the Web, which involves coupling hypertext technologies with a computer network, is a European invention made in Geneva by a European, on the site of the CERN. It is nevertheless in the United States that it first took off. Finally, closer to us, the craze for the Internet at the end of the nineties, and the speculative bubble which followed, also resulted from an error of judgment.

3 Long-term Predictions

In a word, a quick look at the recent history of IST — Information Sciences and Technologies — shows that short and medium-term predictions of, say, five to seven years have very often turned out to be wrong. According to a general principle which says that it is easier to predict that which is close than that which is far, the further away the future, the harder it is to guess what it will hold. Consequently, if the short or medium-term forecasts are incorrect, then the long-term forecasts will be all the more unreliable. But does this mean that it is no longer possible to make predictions in this domain? Reality seems to contradict these intuitions and we can see that many long-term forecasts, over periods of 20, 30 or 50 years, have in fact come true, as the following examples illustrate.

3.1 Moore's Law

Moore's Law stipulates that the speed of processors and their storage capacity double every 18 months. This empirical law was suggested in 1965 by Gordon Moore, co-founder of Intel (manufacturer of microprocessors) and was true for 40 years; the physical principles on which the design of current electronic circuits is based will certainly have to be revised if we wish to continue to make progress at the same pace beyond 2015 or 2020. The fact is that this law remains valid and should remain valid for at least the next 10 years. What we have here is a completely empirical law of prediction which is not based on any rigorous scientific foundation but which has nevertheless been confirmed by long-term experience.

3.2 Alan Turing's Imitation Game

In a famous paper [12] that he wrote in 1950, Alan Turing tried to clarify what it means for a machine "to think". According to him, thought and, more generally, intelligence have nothing to do either with physical appearance, or with voice texture, or even with facial expression. Nor do they have anything to do with consciousness. A machine can be described as an intelligent being if what we observe of its behaviour seems to emanate from an intelligent being. In order to make clear what exactly he understood by the intelligence of machines, Alan Turing imagined a subterfuge called the "imitation game". This is a game for three players: a man, we shall call him A; a woman, B, and an examiner, C, of either sex, it doesn't matter. A, B and C are in three separate rooms so that they cannot perceive each other's voices or physical appearance. In the first step, the examiner, C, has to ask A and B questions in order to distinguish the man from the woman, knowing that the man, A, imitates a woman. At this point there are no computers and no machines in the imitation game. What happens now if we replace the man who imitates the woman by a computer which imitates the man who imitates the woman? According to Alan Turing, a computer could be considered intelligent if it were able to deceive the examiner for as long as the man who imitates a woman. We won't go here into the relevance of this test of intelligence, which has been the subject of much comment, but will focus on what Alan Turing said. According to him, and remember that we were in 1950, we had 50 years to design a computer capable of deceiving an examiner in more than 70

3.3 "2001, A Space Odyssey"

Although Stanley Kubrick's film was first screened in 1968, the script had been written some years earlier, in 1965. If we remove everything which refers to the myth of the all-powerful computer and if we look carefully at the various technologies which were shown there, the film is a pretty faithful reflection of the state of research in American laboratories at that time, i.e. in 1965. Now, 40 years on, we can see that many of these technologies, which were then at the cutting edge of research, are quite normal today, as the following examples show. The spaceship was piloted and regulated by a computer as are many planes, the space shuttle or rockets today. What seemed very innovative at that time is totally commonplace today. Another example, safety is ensured by comparing the results obtained by three computers: if two of them are in conflict with the third, then it is the third which is queried. This principle is used now to ensure the safety of complex systems and we even try, provided it is not too expensive, to get different teams to build several computer programs and activate them in parallel. Remember that in the film the spaceship computer was playing chess and systematically defeated all the adults, whereas at that time, that is to say in 1965, a 10-year-old child beat one of the first chess-playing computers. Since 1997, we all know that a computer is capable of defeating the world chess champion, and this more than once. Finally, the small androids inside the spaceship were able to speak, to understand what the astronauts were saying, and to talk with them. At that time, research on automatic speech recognition and on natural language processing was at a very basic level. Today,

much progress has been made in this field. Now, machines are able to transcribe what we say and to improve their performance by automatically adapting their behaviour to our voice and to our vocabulary. Some people, such as philosophers Dan Sperber[2] or William Crossman [1], think that we are entering a society without writing where there will no longer be any need to write or to type. This means that following the example of the ancients, we shall just have to dictate to computers, which will take the place of the scribes in Antiquity. Let me just point out that, in many laboratories, android robots endowed with speech and vision, and similar in all respects to those of the "2001, A Space Odyssey" spaceship, are today commercially available. Studying science fiction films, if we go beyond the myths and examine what is being done in research laboratories, enables us to make totally satisfactory long-term predictions.

3.4 The Status of Knowledge in a Postmodern Society

At the end of the seventies, a French philosopher, Jean-Francois Lyotard, published a book entitled "The postmodern condition" [2] about the status of knowledge in post-modern societies. This book followed an inquiry ordered by the Canadian government on the long-term consequences of the development of on-line data processing. Without going into the detail of Jean-Francois Lyotard's analysis, let us just remember that the philosopher was examining the consequences of the development of information and communication technologies on the status of knowledge in modern societies. The question was how individuals would be able to read and interpret the knowledge recorded in immense data bases without going through the traditional well-established forms of mediation. More generally, Jean-Francois Lyotard was asking himself how social links would be affected and how the great narratives that legitimated knowledge in a totally decentralized society would be reconstituted. While the words of on-line data processing used at the time seem totally outdated, while the technologies to which Jean-Francois Lyotard referred, in particular the Minitel, are no longer in use, the questions that he asked are totally up to date; they are no longer projections as to the future but are burning issues today. In summary, whereas the first set of examples has shown that short-term forecasts are less and less reliable, the examples here suggest that long-term forecasts can often turn out to be more useful and correct. We are thus in a strange situation where the near future is more difficult to predict than the distant future. Let us now try to understand the origin of this apparent paradox.

4 The Interactive Society

4.1 User-centred Design

The first observation to be made is that the unpredictability of short-term change varies from one area of industrial and economic activity to the next. For instance, the development of electronic components seems relatively predictable in the short term but less predictable in the long term, thus obeying the classic law according to which the further ahead we are looking, the less precise the knowledge we have. So it is with the development of traditional industries, where technological progress does not directly involve social retroactivity. For example, in the production of electricity, in the nuclear, chemical or metal industries, it seems that predictions follow our common intuitions, i.e. the immediate future is more easily predicted than the distant future. Note, too, that it is neither the complexity, nor the profusion of results that makes prediction impossible. In many of the areas mentioned here, a good number of results are both practical and theoretical. And this is not in contradiction with short-term forecasts. If, on the other hand, we take the spheres of activity where the cognitive faculties of the users are involved, or the way in which people appropriate the technologies for themselves, then the classic laws do not apply. Let us take the example of the car industry. In purely technological terms, the development of new cars is dependent on technological progress in the field of both materials and engines. As such, new developments obey the traditional laws of progress and we can therefore predict them with the same degree of (un)certainty with which we predict developments in the domains of physics, materials, mechanical engineering industries or engines. However, in strategic terms, choices depend on social perceptions which are related, for example, to the ecological concerns of a society. At some point these remain totally unpredictable. In the particular case of the car industry, new developments depend mainly on exogenous factors which have nothing to do with technology but everything to do with the desire to limit the number of road accidents, for example, or with people's susceptibility to noise and to atmospheric pollution. Things are very similar in the field of information and communication technologies. This explains why human-machine interfaces, personal computers, personal digital assistants and telecommunication networks have developed in a very unpredictable way. More generally, this means that in a certain number of industries it is no longer enough to devise brilliant plans, as would have been the case with traditional industries at the beginning of the 20th century. Today, we are unable to foresee everything, we need to consider all potential users, and this cannot be done without their actual participation. We thus have to organise consultations and preliminary inquiries, in other words engage the users with the design process in what is known as "User-Centred Design" (cf. [7]). This results in a kind of solidarity between those involved in the first stages of the design, and the notion of users' clubs, which is very common in the field of computing, is a perfect answer. But there is also a practical aspect to this communication strategy: since the designers cannot control all the parameters which will lead to customer satisfaction, they ask for volunteers, who often feel they are part of the chosen few as they can test something for nothing. This is the only way for the designers to identify the weaknesses of the products that they have designed and to adapt them to satisfy the needs of the majority of users.

4.2 The Law of the Second Newcomer

To show just how important user satisfaction is, there is even a law known as the "second newcomer law", which stipulates that on the very advanced technology market the second newcomer possesses a major strategic advantage if he can learn from the failures of his predecessors. Many examples can be given in support of this law. The success of Macintosh in 1984, and the spread of microcomputers with a graphic interface and mouse which followed, had been preceded by numerous unsuccessful attempts which, even though all the ingredients to make the fortune of these machines were there, didn't take off. More recently, some may remember the "Newton", the particularly innovative machine invented by Apple. This computer, with no keyboard but with a touch-sensitive screen and a stylet, prefigured today's pocket computers (which are usually called "palm" computers), UMPC — Ultra Mobile PC — and PC tablets. Although the "Newton", with its graphic interface and automatic handwriting

[2] The Sperber paper in favour of this thesis can be found at $http://www.text-e.org/conf/index.cfm?ConfTextID=12$

recognition, anticipated a whole new generation of machines, it was unsuccessful. Many more illustrations of this law could be given. The point here is that the traditional principle according to which precursors adopt a dominating and dominant place on markets is proving wrong in the field of information and communication technologies. For example, in the field of electronics a huge number of patents have been taken out, which makes it impossible today for anyone to get into the market if he does not already possess a large amount of technological knowledge, because the cost of purchasing all these licenses would be prohibitive. The strategic advantage should thus go, as in traditional industries, to the pioneers who, with their know-how and their intelligence, knew how to be the first ones there. Now, curiously, in the most modern industries where design requires a kind of mutual participation of all, it would seem that the pioneers have a handicap, and it is the second newcomers who defeat the first. The recent history of the development of search engines is a very good illustration of this principle: many financiers decided to invest very early on in the first companies to propose search engines, because they believed they would dominate the market, alone. It is doubtlessly this fear which fed the speculative bubble around the development of the Internet. If we now look closely at the development of these technologies, we see that these fears were ungrounded, on the contrary. To prove my point, we just have to see when some of the most well-known search engines appeared: Incite 1993, Lycos 1994, Yahoo 1994, Altavista 1995, Hotbot 1996, Google 1998, etc. It would seem that the latecomers had the edge over the first newcomers.

4.3 Retroactivity

As soon as we consider this interactive model of design and industrial development, we can no longer think in terms of authoritarian orders simply being transmitted by expert panels with an established and recognized competence. We have to consider all interactions of all possible users during the design and manufacturing processes. To summarize the state of things today, let us consider the comments of a large computer manufacturer who, about twenty years ago, asserted that those who refused to learn the language of technology would be left behind, abandoned at the roadside of modernity. Against this traditional view of progress being imposed on the whole of society, another view would recognise the mutual dependence between designer and user. This is illustrated perfectly by the answer of another computer manufacturer, who said that the manufacturer who does not know how to propose tools which are adapted to the needs and abilities of the users runs the risk of being left by the wayside of economic development and of going bankrupt. Let us now suppose that, in order to understand the causes of the above-mentioned paradox, we try to model progress and development. To do so, we must not take one promoter — the chief engineer — in isolation, nor an interdependent group of people acting jointly — the producers —, but a set of people — including different producers and consumers — interacting with each other, with different competencies and different goals. It follows that the laws of change no longer obey the rules of classical causality, where the effect occurs as a consequence of the cause. Here, it is necessary to consider all possible feedback, which requires using the dynamic system theory where short-term behaviours may be chaotic whereas long-term changes converge. In a word, it is easy to explain this strange above-mentioned phenomenon whereby short-term predictions are more unreliable than long-term ones, through dynamic system theory. Even if there is no pretension of producing a science or a theory of scientific progress, it is at least possible to propose a model which explains the erratic character of predictions.

4.4 A Cybernetic Model

For the sake of clarity, let us take the following example. Suppose we have two technologies, T_1 and T_2, and the knowledge required to design each of the two technologies T_i is made up of two knowledge sources K_i^t and K_i^u. The first, i.e. K_i^t, is just technological and can be acquired either by a single initial investment or, progressively, by repeated investments corresponding to a percentage of the profits. The second source, K_i^u, corresponds to user feedback. It may be empty in the initial state, but not necessarily. It is then possible to express the investment flows between the users, the technologies and the different knowledge sources. Such a very simple dynamic network makes it possible to study different systems of technological innovation. In the first case, user feedback does not play a key role in user satisfaction and therefore in the amount purchased. The determining factor is the technological knowledge. It then appears that the technology T_i which takes advantage of the highest knowledge source K_i^t will provide more user satisfaction and therefore will become dominant. This is especially the case when the two technologies T_1 and T_2 are alternatives, i.e. when users can choose between the two. With time, new knowledge sources K_i^t may appear, which could generate new dominant technologies. In technical terms, the attractors may change if the number of knowledge sources given to a non dominant technology increases. However, this comes at a cost and if the new knowledge sources mean buying the previous one, this will prevent any newcomer from coming on the market. Let us now consider the case where user feedback is required to design a technology T_1. If the initial user feedback is too low, i.e. if K_1^u is almost empty, the design may be wrong, which causes a failure, even if the technological knowledge sources of T_1, i.e. K_i^t, are satisfactory. In the case of two competing technologies, T_1 and T_2, one technology, T_1, for instance, which came first, may turn out to have made bad design choices, while the second can take advantage of the user feedback concerning the first one. The final point is that feedback delay can cause instabilities in the system. Whatever the case, all possible situations and scenarios can be simulated using automata networks, which will show different behaviours corresponding to different systems of technological development, depending on the nature and cost of knowledge required. The first implementation has been carried out using multi-agent architecture, on a NetLogo platform[3]. It shows the different behaviours that have been mentioned here.

5 Plaited Time

By way of conclusion, let us now consider the new structure of time as a result of taking such interactions into consideration. In one way, the time of progress is linear and runs without ever going back on itself, as if it were an arrow flying forward, pursuing its course without knowing if and when this will end. This linear time of perpetually renewed modernity is in opposition to traditional cyclic time, in which the future is never other than a return to the past, which means that time is eminently predictable since nothing really new can happen. In the present case, it seems obvious that time cannot be conceived of as being cyclic, because technological and industrial developments impose endless renewal: nothing is today as it was yesterday. The contemporary imperative, which orders us to be modern, is proof of the singular novelty of modernity. As a consequence, the present cannot be considered as the return to a former present. Does this mean

[3] NetLogo is a free multi-agent modelling environment that is available at http://ccl.northwestern.edu/netlogo/

that current time must be seen as being strictly linear? It is undoubtedly a time of progress, which accumulates results and which thus every day opens up new perspectives. In this respect, we could indeed be tempted to see it as being linear. Nevertheless, reducing time to a straight line would be misleading. After all, as we have just seen in this paper, present time surprises us, its progress sometimes takes on a chaotic look. It goes back to the past. Paths which had been trodden in the course of investigation, then abandoned, reappear and are successful. Some strands of change are divided and subdivided to such an extent that the thread of time seems to be forked and twisted rather than simply linear. But in spite of these twists and forks, the long-term predictions are relatively stable since, quite often, those that came too soon are rejected for a while, before returning to the front of the stage. The structure of time is therefore not really ordered like shelf space is, but is somewhat tangled. In other words, at any given point several alternatives can be envisaged, some of which exercise the most brilliant minds, while others seem to be in retreat, hidden from the general public. Then, from time to time, what seemed hidden reappears and what has gradually been emerging into view disappears. Time is thus a tangled hank, a plait of hair, its strands scattering and even moving out of sight before reappearing under a new light, then hiding anew. It is in this sense that we can speak about the time of contemporary modernity as plaited time.

ACKNOWLEDGEMENTS

I would like to thank the referees for their helpful comments which enabled me to improve this paper.

REFERENCES

[1] W. Crossman, *VIVO [Voice-InVoice-Out]: The Coming Age of Talking Computers*, Regent Press, 2004.
[2] J.-F. Lyotard, *La condition post-moderne*, Editions de Minuit, Paris, 1979.
[3] W. Mc Culloch and W. Pitts, 'A logical calculus of the ideas immanent in neuron activity', *Bulletin of Mathematical Biophysics*, (1943).
[4] J. McCarthy, M. Minsky, N. Rochester, and C. Shannon, 'A proposal for the dartmouth summer research project on artificial intelligence: August 31, 1955', *AI Magazine*, (December 2006).
[5] M. Minsky and S. Papert, *Perceptrons*, MIT Press, Cambridge, MA, 1969.
[6] E. Morin and Kern A.B., *Homeland Earth : A Manifesto for the New Millennium (Advances in Systems Theory, Complexity and the Human Sciences)*, Hampton Press, 1999.
[7] D. Norman, *The Psychology of Everyday Things*, Basic Book, 1988.
[8] J. Pierce and J. Carroll, *Language and Machines Computers in Translation and Linguistics. ALPAC report*, National Academy of Sciences, National Research Council, Washington, DC, 1966.
[9] F. Rosenblatt, 'The perceptron: A probabilistic model for information storage and organization in the brain', *Psychological Review*, **65**, 386–408, (1958).
[10] A. Rosenblueth, N. Wiener, and J. Bigelow, 'Behavior, purpose and teleology', *Philosophy of Science*, **10**, 18–24, (1943).
[11] H. Simon and A. Newell, 'Heuristic problem solving: the next advance in operations research', *Operations Research*, **6**, 1–10, (1958).
[12] A. Turing, 'Computing machinery and intelligence', *Mind*, **59**, 433–460, (1950).

Substitution for Fraenkel-Mostowski foundations

Murdoch J. Gabbay[1] and **Michael J. Gabbay**[2]

Abstract. A fundamental and unanalysed logical concept is *substitution*. This seemingly innocuous operation — substituting a variable for a term or valuating a variable to an element of a domain — is hard to characterise other than by concrete constructions. It is widely viewed as a technicality to be dispensed with on the way to studying other things. Discussions of computer science foundations, and of the philosophy of logic, have largely ignored it.

We show that Fraenkel-Mostowski set theory gives a model of variables and substitution as constructions on sets. Thus models of variables and substitution are exhibited as constructions in a foundational universe, just like models of arithmetic (the ordinals) and other mathematical entities. The door is open for classes of denotations in which variables, substitution, and evaluations are constructed directly in sets and studied independently of syntax, in ways which would previously have not been possible.

1 Introduction

Computer science evolved out of the study of logic and the foundations of mathematics of the late 19th and early 20th centuries. Part of the motivation for that study was to devise a framework to explain mathematical knowledge, as can clearly by seen in the work of Gottlob Frege [10, 11]. Frege's development of predicate calculus was, amongst other things, intended to explain the content of statements about 'an arbitrary number' without positing some special entity that is an arbitrary-number. In terms of the modern predicate calculus, the solution was that a statement about an arbitrary number has the form of an open or universally quantified sentence. So: the content of $A(x)$ or $\forall x.A(x)$ is given by the contents of $A(x)[x/t]$ for all t that denote elements of a certain domain.

This approach leaves unexplained the phenomenon of the substitution of x for a term. Substitution is not trivial: substitutions may occur within other substitutions and when a substitution is carried out, variables must be renamed to ensure that no unwanted bindings result. Therefore the explanation of knowledge of generic mathematical statements in terms of substitutions just changes the issue to the explanation of knowledge about substitution. We might ask again what exactly we know when we know that $A(x)[x/t]$ for any t.

In fact, attempts to formalise the theory of substitution show that the 'explanation' in terms of substitutions just makes matters worse. The complexity of the knowledge needing an explanation has increased, if the content of $A(x)$ has been given in terms of the more complex $A(x)[x/t]$. "What is the content of the arbitrary t in $A(x)[x/t]$?" we may ask — an answer in terms of the even more complex $A(x)[x/y][y/t]$ is of no help.

We find no relief in replacing talk of substitution with talk of valuations: this merely translates the problem into a different language; the content of $\forall x.A(x)$ is given in terms of valuations $A(x)\langle x{\mapsto}d\rangle$ for all d — but what is a valuation, and what do we know when we know the generic statement that, say $A(x)\langle x{\mapsto}d\rangle = \top$ for all valuations on x?

This problem is not confined to the philosophical question of the content of an open or universally quantified statement. All formal languages used to express functions and computations, and reasoning about functions and computations, refer to substitution of variables for terms or to the resolution of a variable to a value. An account of the content of substitution and valuation would therefore shed light on functions and computation.

But are we misguided, or asking for too much, when we ask for an explanation of substitution? After all, the syntax of a formal language is impossible to formulate without using schematic variables and substitution. Research into computer science cannot get off the ground without, at least, some recourse to formal syntax. This suggests that we must be satisfied to take substitution as an unanalysable primitive — we must either take substitution as a purely formal syntactic manipulation, or accept that the only possible explanation of substitution on the syntax of one formal language must be in terms of substitution in another 'meta' formal language (the so-called Higher-Order Abstract Syntax approach [19]). In either case, we must give up on trying to provide a foundational theory to account for substitution independently of formal syntax.

In this paper we shall show that, on the contrary, there is an independent foundation that can interpret the action of substitution and valuation. This foundation is called Fraenkel-Mostowski set theory.[3]

Fraenkel-Mostowski set theory [6, 22] (**FM** set theory) was originally developed to prove the independence of the axiom of Choice from the other axioms of Zermelo-Fraenkel set theory. It was rediscovered and used by the first author and Pitts to model abstract syntax with binding [16]. An advantage of modelling syntax in a model of FM set theory is that datatypes of syntax quotiented by α-equivalence can be modelled inductively (rather than as *quotients* by α-equivalence of syntax-without-binding). This is because FM set

[1] http://www.gabbay.org.uk

[2] michael.gabbay(at)kcl.ac.uk, Michael Gabbay gratefully acknowledges the support of the British Academy under grant PDF/2006/509.

[3] We know of no set-theoretic foundational account of substitution in the literature, besides this paper. However, there have been many attempts to axiomatise the properties that such an account should have.
Fine [9] has axiomatised 'arbitrary objects', especially investigation of *dependency between arbitrary objects*; the intuition is that both x and $2 * x$ are arbitrary objects, but they are correlated. It remains to be seen whether a model of FM set theory can be considered as a model of Fine's axioms.
Aczel's 'generalised set theory' [3] and 'universes with parameters' [4] model variable symbols (Aczel calls them *parameters*) as atoms in a ZFA-like set theory. The resemblance ends there; Aczel imposes all the structure he needs as explicit axioms on names, and the substitution action is not capture-avoiding, which is one of the most difficult technical aspects of the work in this paper. The application is also quite different; Aczel investigates coinductive structures and non-wellfounded set theory [2] as a semantics for behaviour.

theory delivers a model of variable symbols and α-abstraction [16] — these feature in this paper as atoms (Subsection 2) and atoms-abstraction (Subsection 2.4).

Unlike HOAS [19] there is no problem with 'exotic terms'; also more functions, such as α-inequality, may be expressed; finally, and there is no need in FM for levels of carefully-constrained meta-language. Unlike de Bruijn indexes [8] the reasoning and programming principles of syntax-with-binding in FM are natural and correspond very closely to informal practice. Seven years of research, culminating in an implementation of these ideas in Isabelle [23] have demonstrated the practical potential of this technique.

What makes this all take off is that the model of variable symbols and α-abstraction provided by FM set theory is applicable to all sets, including those modelling functions, predicates, domains, games, and so on. They can be applied to denotations other than sets modelling syntax. Since the introduction of these ideas [16] there now exist programming languages [21, 7], logics [20, 14], models of storage [5], and semantics of references using game theory [1] — research continues and the work all uses the model of variable symbols and α-abstraction which emerges from FM set theory.

However we usually are interested in variable symbols and α-abstraction because we want a *capture-avoiding substitution action*. In this paper we demonstrate that the variable symbols in models of FM set theory admit a substitution action defined as an operation between arbitrary sets. We also show that this substitution action avoids capture with α-abstraction. In short, any model of FM set theory is *also* a model of something that looks like 'substitution' in formal syntax, but which is valid for all sets.

We envisage denotations using FM set theory in which variables and open terms are explained directly as sets — without needing valuations — and substitution in syntax is explained directly as substitution on sets.

2 Fraenkel-Mostowski set theory

2.1 Axioms, permutations, equivariance

The language of FM set theory is first-order logic with binary predicates $=$ (set equality) and $\in$ (set membership) — like the language of ZF set theory — *and* one constant symbol $\mathbb{A}$ for 'the set of atoms'.

Definition 1. *The axioms of FM set theory are given in Figure 1.*

In Figure 1 we use standard definitional extensions of the language of sets. $\mathcal{P}_{fin}(\mathbb{A})$ is the finite powerset of $\mathbb{A}$ (the set of finite subsets of $\mathbb{A}$). 'S supports x' is described in Definition 4. The standard cumulative hierarchy model of these axioms is described in Remark 8.

We will use some notational conventions in the rest of this paper:

- An **atom** is a set member of $\mathbb{A}$ (the set of atoms).
- *The permutative convention:* $a, b, c, \ldots$ range over *distinct* atoms unless stated otherwise.
- A, B, C, S, T range over sets of atoms. For example $A \subseteq \mathbb{A}$.
- X, Y, Z, U, V range over elements that are not atoms and may be empty. For example X might equal $\emptyset$ or $\{a, \emptyset\}$, but X cannot equal a.
- x, y, z, u, v range over arbitrary elements.

Remark 2. An atom $a \in \mathbb{A}$ is 'empty' ($\forall x.x \notin a$) but not equal to $\emptyset$. $(\mathbf{Extensionality})$ is weakened so that an empty element is equal to $\emptyset$, or is an atom.

Note that $(\mathbf{AtmInf})$ insists that there are infinitely many atoms.

2.2 Atoms, equivariance and support

Write $(a\ b)$ for the **swapping** function from atoms to atoms:

$$(a\ b)(a) = b \qquad (a\ b)(b) = a \qquad (a\ b)(c) = c.$$

By our permutative convention, a, b, and c are distinct.

Let π range over functions generated by composing finitely many swappings, call these functions **permutations**. Write $\circ$ for functional composition and π^{-1} for the inverse of π, which is also a permutation. The action of permutations extends to all sets by ϵ-induction [18]:

$$\pi X = \{\pi x \mid x \in X\}.$$

Let $\phi(x_1, \ldots, x_n)$ range over predicates in the language of FM set theory that mention variables in $x_1, \ldots, x_n$. An n-rary function $F(x_1, \ldots, x_n)$ can be expressed by an $n{+}1$-ary predicate $\phi_F(x_1, \ldots, x_n, z)$ such that for each $x_1, \ldots, x_n$ there is a unique z making ϕ_F true. Then **equivariance** is the following two properties:

Theorem 3. *$\phi(x_1, \ldots, x_n) \Leftrightarrow \phi(\pi x_1, \ldots, \pi x_n)$, and $\pi(F(x_1, \ldots, x_n)) = F(\pi x_1, \ldots, \pi x_n)$ always hold.*

Proof. The first part is by an easy induction on the syntax of ϕ. We consider just one case: $x \in y$ implies $\pi x \in \pi y$ follows directly from the fact that $\pi y = \{\pi y' \mid y' \in y\}$. The reverse implication uses π^{-1}.

The second part follows using the standard encoding of an n-ary function as an $n{+}1$-ary predicate. □

Equivariance (Theorem 3) holds because atoms have no internal set structure. It is a useful source of one-line proofs [14, 12]; we shall exploit that in this paper. Equivariance is also a sense in which atoms are 'abstract': if we pick some sets, containing some specific atoms, and prove a property of them, then that property is as true of the sets with the atoms permuted; the identity of atoms only matters up to permutations.

2.3 Support

Definition 4. *If $S \subseteq \mathbb{A}$ write $\mathsf{fix}(S) = \{\pi \mid \forall a \in A.\pi(a) = a\}$.*
*Say that $S \subseteq \mathbb{A}$ **supports** x when $\forall \pi \in \mathsf{fix}(A).\pi x = x$.*
*Define $supp(x)$ the **support** of x by:*

$$supp(x) = \bigcap\{S \mid S \text{ is finite, } S \text{ supports } x\}.$$

$supp(x)$ always exists in FM set theory because $(\mathbf{Fresh})$ insists that a finite S supporting x exists. Write $a\#x$ when $a \notin supp(x)$. Read this 'a is **fresh for** x'. We may write $a\#t_1, t_2$ for '$a\#t_1$ and $a\#t_2$', and so on.

Remark 5. For example:

- $supp(\emptyset) = \emptyset$. $\quad \pi\emptyset = \emptyset$ for all $\pi \in \mathsf{fix}(\emptyset)$.
- $supp(\mathbb{A}) = \emptyset$.
  $\pi\{a, b, c, \ldots\} = \{\pi(a), \pi(b), \pi(c), \ldots\} = \{a, b, c, d, \ldots\}$ for all $\pi \in \mathsf{fix}(\emptyset)$.
- $supp(a) = \{a\}$. $\quad \pi(a) = a$ for all $\pi \in \mathsf{fix}(\{a\})$.
- $supp(\{a\}) = \{a\}$. $\quad \pi(\{a\}) = \{a\}$ for all $\pi \in \mathsf{fix}(\{a\})$.
- $supp(\mathbb{A} \setminus \{a\}) = \{a\}$.
  $\pi\{b, c, d, \ldots\} = \{\pi(b), \pi(c), \pi(d), \ldots\} = \{b, c, d, \ldots\}$ for all $\pi \in \mathsf{fix}(\{a\})$.
- $supp(\{a, b\}) = \{a, b\}$. $\quad \pi\{a, b\}{=}\{a, b\}$ for all $\pi{\in}\mathsf{fix}(\{a, b\})$.

$$(\textbf{Sets})\quad \forall x.(\exists y.y \in x) \Rightarrow x \notin \mathbb{A} \qquad (\textbf{Extensionality})\quad \forall x.x \notin \mathbb{A} \Rightarrow x = \{z \mid z \in x\}$$
$$(\textbf{Comprehension})\quad \forall x.\exists y.y \notin \mathbb{A} \wedge y = \{z \in x \mid \phi(z)\} \quad (y \text{ not free in } \phi) \qquad (\epsilon\textbf{-Induction})\quad \big(\forall x.(\forall y \in x.\phi(y)) \Rightarrow \phi(x)\big) \Rightarrow \forall x.\phi(x)$$
$$(\textbf{Replacement})\quad \forall x.\exists z.z \notin \mathbb{A} \wedge z = \{F(y) \mid y \in x\} \qquad (\textbf{Pairset})\quad \forall x, y.\exists z.z = \{x, y\}$$
$$(\textbf{Union})\quad \forall x.\exists z.z \notin \mathbb{A} \wedge z = \{y \mid \exists y'.(y \in y' \wedge y' \in x)\} \qquad (\textbf{Powerset})\quad \forall x.\exists z.z = \{y \mid y \subseteq x\}$$
$$(\textbf{Infinity})\quad \exists x.\emptyset \in x \wedge \forall y.y \in x \Rightarrow y \cup \{y\} \in x \qquad (\textbf{AtmInf})\quad \mathbb{A} \notin \mathcal{P}_{fin}(\mathbb{A}) \qquad (\textbf{Fresh})\quad \forall x.\exists S \in \mathcal{P}_{fin}(\mathbb{A}).S \text{ supports } x$$

Figure 1. Axioms of FM set theory

- $supp(\mathbb{A} \setminus \{a, b\}) = \{a, b\}$.
  $\pi\{c, d, e, \ldots\} = \{\pi(c), \pi(d), \pi(e), \ldots\} = \{c, d, e, \ldots\}$ for all $\pi \in \mathsf{fix}(\{a, b\})$.
- $supp(\{a, \{a\}, \{c\}, \{d\}, \ldots\}) = \{a, b\}$.

$$\begin{aligned}\pi(\{a,\{a\},\{c\},\{d\},\ldots\}) &= \{\pi(a),\{\pi(a)\},\{\pi(c)\},\{\pi(d)\},\ldots\}\\ &= \{a,\{a\},\{c\},\{d\},\ldots\}\end{aligned}$$

  provided that $\pi \in \mathsf{fix}(\{a, b\})$.

Remark 6. Ideas from syntax match ideas from FM sets as follows: *variable symbols* matches *atoms* and *free variables* matches *support*. Of course, it is possible to take the complement of a set, but not possible to take the complement of a syntax tree. It is therefore important to understand that sets are more general than syntax, and in particular that $a \notin X$ and $a\#X$ are *not* the same thing. $supp(x)$ measures how 'conspicuous' a is in x, either by its set-membership or *lack* of set membership. For example:

$$a \in \mathbb{A} \text{ and } a\#\mathbb{A} \qquad a \notin \emptyset \text{ and } a\#\emptyset \qquad a \notin a \text{ and } a \in supp(a)$$
$$a \in \{a\} \text{ and } a \in supp(\{a\}) \qquad a \notin \mathbb{A}\setminus\{a\} \text{ and } a \in supp(\mathbb{A}\setminus\{a\})$$

Remark 7. Not every collection has finite support. $\{a, c, e, g, \ldots\}$ (the set of 'every other atom') is not finitely supported, and is excluded from the cumulative hierarchy model of Remark 8 below. There is no finite $S \subseteq \mathbb{A}$ such that if $\pi \in \mathsf{fix}(S)$ then $\pi\{a, c, e, g, \ldots\} = \{a, c, e, g, \ldots\}$.

Remark 8. FM is a theory in first-order logic. As is often the case, we have a clear intuition in mind for a standard model; the *cumulative hierarchy* model is the collection $\mathcal{U}$ defined as follows:

$$\begin{aligned}\mathcal{U}_0 &= \mathbb{A}\\ \mathcal{U}_{i+1} &= \mathcal{U}_i \cup \{X \subseteq \mathcal{U}_i \mid X \text{ has a finite supporting set}\}\end{aligned}$$

Then $\mathcal{U} = \bigcup_i \mathcal{U}_i$. The reader can imagine all our constructions taking place in this model and no harm will come of it.

Theorem 9. *If S and T support x and are finite, then so does $S \cap T$. As a corollary, $supp(x)$ is the unique smallest set supporting x.*

Proof. The corollary follows by calculations and by (**Fresh**).

Suppose κ fixes $S \cap T$ pointwise. We must show $\kappa x = x$.

Write K for $\{a \mid \kappa(a) \neq a\}$. Choose an injection ι of $T \setminus S$ into $\mathbb{A} \setminus (S \cup T \cup K)$ (we can say 'ι freshens $T \setminus S$'). Let $\pi(a) = \iota(a)$ and $\pi(\iota(a)) = a$ for $a \in T \setminus S$, and $\pi(a) = a$ otherwise. Note that $\pi \circ \pi = \mathbf{Id}$, so $\pi = \pi^{-1}$. π fixes S pointwise so $\pi x = x$. Also $\pi \circ \kappa \circ \pi$ fixes T pointwise so $(\pi \circ \kappa \circ \pi)x = x$. We apply π to both sides and simplify and conclude that $\kappa x = x$ as required. □

Theorem 9 says πx depends *only* on the values of π on atoms in $supp(x)$. Support goes back to Fraenkel and Mostowski [17, Chapter 4]; applications in computer science followed later [16, 12].

Theorem 10. *S supports x if and only if πS supports πx. As a corollary, $\pi supp(x) = supp(\pi x)$.*

Proof. From Theorem 3. □

A calculation cannot 'create support' not in its inputs:

Theorem 11. $supp(F(x_1, \ldots, x_n)) \subseteq supp(x_1) \cup \cdots \cup supp(x_n)$.

Proof. If $\pi \in \mathsf{fix}(\bigcup supp(x_i))$ then $\pi \in \bigcap \mathsf{fix}(x_i)$. By Theorem 3 $\pi F(x_1, \ldots, x_n) = F(\pi x_1, \ldots, \pi x_n)$. The result follows. □

2.4 α-abstraction in models of FM set theory

Substitution is interesting and hard to characterise because of its interaction with α-*equivalence*, itself deceptively complex. For example, we distinguish x and y in Px and Py but not in $\forall x.Px$ and $\forall y.Py$. We now show how to α-abstract an atom a in a set x. With sets, it is standard to abstract by taking an equivalence class. For example the concept 'even number' can be modelled as the collection of even numbers. Intuitively Definition 12 defines an equivalence class resulting from renaming atoms not in A, and thus α-abstracts over atoms in $supp(x)\setminus A$. This is then exploited in Definition 16.

Definition 12. *Suppose $A \subseteq \mathbb{A}$. Write*

$$u\|_A \quad \textit{for} \quad \{\pi u \mid \pi \in \mathsf{fix}(A)\}.$$

(Recall that $\mathsf{fix}(A) = \{\pi \mid \forall a \in A.\pi(a) = a\}$.) For example:

$$\{a\}\|_\emptyset = \{a,\ b,\ c,\ d,\ e,\ f,\ \ldots\} \qquad \{a\}\|_{\{a\}} = \{a\}$$
$$\{b\}\|_{\{a\}} = \{b,\ c,\ d,\ e,\ f,\ \ldots\}$$
$$\{a, b\}\|_{\{a,c\}} = \{\{a, b\},\ \{a, d\},\ \{a, e\},\ \{a, f\},\ \ldots\}$$

Since $\mathsf{fix}(A)$ is a group we have:

Lemma 13. *If $\pi \in \mathsf{fix}(A)$ then $u\|_A = (\pi u)\|_A$.*

In words: $u\|_A$ is an equivalence class of sets which are equal 'up to renaming atoms not in A'.

Theorem 14. *Suppose A is a finite set of atoms. Then:*

- *If $supp(u) \subseteq A$ then $supp(u\|_A) = supp(u)$.*
- *$supp(u\|_A) \subseteq A$ always.*

As a corollary, if $supp(u) \setminus A \neq \emptyset$ then $supp(u\|_A) = A$.

Proof.
- If $supp(u) \subseteq A$ then $u\|_A = \{u\}$. For example, $supp(a\|_{\{a\}}) = supp(\{a\}) = \{a\}$ and $supp(a) = a$.
- If $supp(u) \subseteq A$ then we use the first part. If there is some $a \in supp(u) \setminus A$ then the result follows by an easy calculation illustrated by the following example:

$$a\|_{\{b\}} = \{a,\ c,\ d,\ e,\ f,\ \ldots\} = \mathbb{A} \setminus \{b\}.$$

The corollary follows. □

Definition 15. *Write* (x,y) *for* $\{\{x\},\{x,y\}\}$ *(a set implementation of ordered pairs [18]).*

Definition 16. *Let* **atoms abstraction** *be* $[c]z = (c,z)\|_{supp(z)\setminus\{c\}}$.

Intuitively $[c]z$ is an α-equivalence class of (c,z) where c is abstracted, i.e. where we read (c,z) like '$\lambda c.z$' or '$\forall c.z$':

$$\begin{aligned}
[a]a &= \{(a,a),\ (b,b),\ (c,c),\ (d,d),\ (e,e),\ (f,f),\ \ldots\}\\
[a]\{a,b\} &= \{(a,\{a,b\}),\ (c,\{c,b\}),\ (d,\{d,b\}),\ (e,\{e,b\}),\ \ldots\}\\
[a](\mathbb{A}\setminus\{a\}) &= \{(a,\mathbb{A}\setminus\{a\}),\ (b,\mathbb{A}\setminus\{b\}),\ (c,\mathbb{A}\setminus\{c\}),\ \ldots\}\\
[a](\mathbb{A}\setminus\{a,b\}) &= \{(a,\mathbb{A}\setminus\{a,b\}),\ (c,\mathbb{A}\setminus\{c,b\}),\ (d,\mathbb{A}\setminus\{d,b\}),\ \ldots\}
\end{aligned}$$

Write U_{ab} for $\{a\}\cup\{\{a\},\{c\},\{d\},\{e\},\ldots\}$ for any a,b. Then:

$$\begin{aligned}
[a]U_{ab} &= \{(a,U_{ab}),(c,U_{cb}),(d,U_{db})(e,U_{eb}),\ldots\}\\
[c]U_{ab} &= \{(c,U_{ab}),(d,U_{ab}),(e,U_{ab}),\ldots\}
\end{aligned}$$

We can read '$[a]x$' as the binding action of '$\lambda a.x$' or '$\forall a.x$', and the sets above coresond with α-equivalence classes of FM sets. There is no *a priori* notion of λ-abstraction or universal quantification in $[a]x$; this is just α-abstraction, on FM sets.

Definition 16 agrees with the definition of $[c]z$ from [16]:

Lemma 17. $[c]z = \{(x,(x\ c)z) \mid x\in\mathbb{A},\ x\neq c,\ x\#z\}\cup\{(c,z)\}$.

2.5 Further properties of support, finite sets, and α-abstraction

Lemma 18. *1.* $supp(X) = \bigcup\{supp(x)\mid x\in X\}$ *if* X *is finite.*
2. $supp(\{x\}) = supp(x)$ *and if* $A\subseteq\mathbb{A}$ *is finite then* $supp(A) = A$.
3. $supp((x,y)) = supp(x)\cup supp(y)$.

Proof. If X is finite then $supp(X)\subseteq\bigcup\{supp(x)\mid x\in X\}$ follows by Theorem 11.

Now suppose $a\in supp(x)$ for some $x\in X$. Choose some b such that $b\#X$ and $b\#x'$ for every $x'\in X$. By Theorem 10 $supp((b\ a)x) = (b\ a)supp(x)$. Since X has no element y such that $b\in supp(y)$, we know that $(b\ a)X\neq X$ and by Theorem 9 it must be that $a\in supp(X)$.

The second part is immediate; the third is by Definition 15. □

Theorem 19. $supp([c]z) = supp(z)\setminus\{c\}$.

Proof. By part 3 of Lemma 18 $supp((c,z)) = supp(z)\cup\{c\}$. By definition $[c]z = (c,z)\|_{supp(z)\setminus\{c\}}$. The result follows by Theorem 14. □

Thus we expect $(a\ d)[a]\{a,b\} = [a]\{a,b\}$:

$$\begin{aligned}
[a]\{a,b\} &= \{(a,\{a,b\}),\ (c,\{c,b\}),\ (d,\{d,b\}),\ (e,\{e,b\}),\ \ldots\}\\
(a\ d)[a]\{a,b\} &= \{(d,\{d,b\}),\ (c,\{c,b\}),\ (a,\{a,b\}),\ (e,\{e,b\}),\ \ldots\}
\end{aligned}$$

Lemma 20. $supp(X)\subseteq\bigcup\{supp(x)\mid x\in X\}$ *need not necessarily hold if* X *is not finite.*

Proof. It suffices to give a counterexample; we give two:
$supp(\mathbb{A}) = \emptyset$ but $\bigcup\{supp(a)\mid a\in\mathbb{A}\} = \mathbb{A}$.
$supp(\mathbb{A}\setminus\{c\})=\{c\}$ but $\bigcup\{supp(a)\mid a\in\mathbb{A}\wedge a\neq c\} = \mathbb{A}\setminus\{c\}$. □

3 The substitution action

We now turn to defining an operation on sets that matches the syntactic operation of substitution. It must interact correctly with the α-abstraction of Definition 16.

Recall that a,b,c range over distinct atoms, A,B,C,S,T range over sets of atoms, x,y,z,u,v range over all elements, and X,Y,Z,U,V range over elements that are not atoms.

3.1 Axioms, naïve substitution action

Definition 21. *A* **substitution action** *on FM set theory is a function* $z[a{\mapsto}x]$ *expressed in the language of FM set theory taking an element* z*, an atom* a*, and an element* x*, and returning an element which we write as* $z[a{\mapsto}x]$*, satisfying:*

$$\begin{aligned}
(\alpha)\quad & b\#z\Rightarrow z[a{\mapsto}x] = ((b\ a)z)[b{\mapsto}x]\\
(\#{\mapsto})\quad & a\#z\Rightarrow z[a{\mapsto}x] = z\\
(\mathbf{var}{\mapsto})\quad & a[a{\mapsto}x] = x\\
(\mathbf{id}{\mapsto})\quad & z[a{\mapsto}a] = z\\
(\mathbf{abs}{\mapsto})\quad & c\#x\Rightarrow([c]z)[a{\mapsto}x] = [c](z[a{\mapsto}x])
\end{aligned}$$

If we read $a\#x$ as 'a is not free in x' and $z[a{\mapsto}x]$ as 'substitute x for a in z' then, clearly, the axioms of Definition 21 are sound for the standard syntactic model. In [13] they are also proved complete.[4]

FM set theory has notions of 'name' and 'free in', and 'abstraction'. We can therefore try to build a function which models 'capture-avoiding substitution' in the sense made precise by the axioms of Definition 21.

Definition 22 is probably what we might first consider:

Definition 22. *Define the* **naïve** *substitution action by*

$$a[a{\mapsto}x]_n = x\quad b[a{\mapsto}x]_n = b\quad Z[a{\mapsto}x]_n = \{z[a{\mapsto}x]_n\mid z\in Z\}.$$

Write $0=\emptyset$ and $i+1 = i\cup\{i\}$, and write $\mathbb{N} = \{0,1,2,3,\ldots\}$.

Lemma 23. *Naïve substitution does not satisfy* (α)*,* $(\#{\mapsto})$*, or* $(\mathbf{abs}{\mapsto})$*, and so is not a substitution action in the sense of Definition 21.*

Proof. It suffices to give counterexamples. We do this for $(\#{\mapsto})$ and $(\mathbf{abs}{\mapsto})$. We expect that $\mathbb{A}[a{\mapsto}1]_n = \mathbb{A}$ since $a\#\mathbb{A}$. We also expect that $([c]a)[a{\mapsto}1]_n = [c](a[a{\mapsto}1])$ since $c\#1$. But:

$$\begin{aligned}
\mathbb{A}[a{\mapsto}1]_n &= (\mathbb{A}\setminus\{a\})\cup\{1\}\\
([c]a)[a{\mapsto}1]_n &= \{(b,a),(c,a),(d,a),(e,a),\ldots\}[a{\mapsto}1]_n\\
&= \{(b,1),(c,1),(d,1),(e,1),\ldots\}\\
[c](a[a{\mapsto}1]_n) &= [c]1\\
&= \{(a,1),(b,1),(c,1),(d,1),(e,1),\ldots\}.
\end{aligned}$$
□

The naïve substitution action does not take into account that substitutions should be capture avoiding and does not interact property with the FM treatment of abstraction. We need a more subtle substitution action that 'unpacks' an FM set to discern the 'free' atoms and equate the 'bound' atoms. The basic units of such an unpacking are the *planes* defined in Definition 24.

3.2 The planes of a set

Definition 24. *If* $A\subseteq\mathbb{A}$ *is finite call* (u,A) *a* **plane** *in* Z *when*

- $u\|_A\subseteq Z$ *and* $A\subseteq supp(Z)$*, and*
- $u\|_A$ *is maximal in that for all* $u'\|_{A'}\subseteq Z$ *where* $A'\subseteq supp(Z)$*,*

$$u\|_A\subseteq u'\|_{A'}\quad\textit{implies}\quad u'\|_{A'} = u\|_A.$$

[4] An equivariance rule from [13] is omitted here because it is guaranteed by Theorem 3. Instead of $(\mathbf{id}{\mapsto})$ we use a rule $(\mathbf{ren}{\mapsto})$ in [13]. A proof that the two formulations are equivalent is not hard (and was observed by an anonymous referee of [13]). The proof is included in a recent work pending publication.

Write $\mathsf{plane}(Z)$ *for the collection of planes in* Z.

(u, A) is a plane in Z when A is a least subset of $supp(Z)$ such that $u\|_A \subseteq Z$. For example:

1. $(a, \{a\}) \in \mathsf{plane}(\{a\})$ and $a\|_{\{a\}} = \{a\} \subseteq \{a\}$.
2. $(a, \{\}) \notin \mathsf{plane}(\{a\})$ because $a\|_{\{\}} = \mathbb{A} \not\subseteq \{a\}$.
3. $(a, \{a\}) \notin \mathsf{plane}(\mathbb{A})$ because $\{a\} \not\subseteq supp(\mathbb{A}) = \emptyset$.
4. $(a, \{a, b\}) \notin \mathsf{plane}(\{a\})$ because $a\|_{\{a,b\}} = \{a\} = a\|_{\{a\}}$ and $\{a\} \not\subseteq \{a, b\}$.
5. $(c, \{a\}) \in \mathsf{plane}(\mathbb{A}\backslash\{a\})$ and $c\|_{\{a\}} = \mathbb{A}\backslash\{a\} \subseteq \mathbb{A}\backslash\{a\}$.
6. $(a, \{\}) \in \mathsf{plane}(\mathbb{A})$ and $a\|_{\{\}} = \mathbb{A} \subseteq \mathbb{A}$.
7. $((c, a), \{a\}) \in \mathsf{plane}([c]a)$ and $(c, a)\|_{\{a\}} = \{(x, a) \mid x \neq a\} = [c]a \subseteq [c]a$.
8. $\mathsf{plane}(\{a\} \cup \{\{a\}, \{c\}, \{d\}, \ldots\}) = \{(a, \{a\})\} \cup \{(\{x\}, \{b\}) \mid x \in \mathbb{A},\ x \neq b\}$.
   $a\|_{\{a\}} = \{a\} \subseteq \{a\} \cup \{\{a\}, \{c\}, \{d\}, \ldots\}$.
   $\{x\}\|_{\{b\}} = \{\{a\}, \{c\}, \{d\}, \ldots\} \subseteq \{a\} \cup \{\{a\}, \{c\}, \{d\}, \ldots\}$.

Definition 25. *If* $S \subseteq \mathbb{A}$ *is finite then define*

$$\mathsf{plane}_S(Z) = \{(u, A) \in \mathsf{plane}(Z) \mid supp(u) \cap S \subseteq supp(u) \cap A\}.$$

We should think of $\mathsf{plane}_S(Z)$ as the planes (u, A) in Z such that $supp(u)$ 'avoids name-clashes' with S. For example

$$(a, \{\}) \in \mathsf{plane}_{\{\}}(\mathbb{A}) \quad \text{but} \quad (a, \{\}) \notin \mathsf{plane}_{\{a\}}(\mathbb{A}) \quad \text{and}$$
$$(c, \{a\}) \in \mathsf{plane}_{\{b\}}(\mathbb{A}\backslash\{a\}) \quad \text{but} \quad (b, \{a\}) \notin \mathsf{plane}_{\{b\}}(\mathbb{A}\backslash\{a\}).$$

The planes of Z 'cover' Z in the following sense:

Lemma 26. *If* $S \subseteq \mathbb{A}$ *is finite then*

$$\bigcup\{u\|_A \mid (u, A) \in \mathsf{plane}_S(Z)\} = Z.$$

As a corollary taking $S = \emptyset$, $\bigcup\{u\|_A \mid (u, A) \in \mathsf{plane}(Z)\} = Z$.

Proof. We prove two set inclusions: The left-to-right inclusion is by construction. For the right-to-left inclusion, choose any $u \in Z$. Let $B = \{b_1, \ldots, b_k\}$ be equal to $supp(u) \setminus A$ and let $B' = \{b'_1, \ldots, b'_k\}$ be some set of entirely fresh atoms (so disjoint from $supp(u)$, A, S, and $supp(Z)$). Let $\pi = (b_1\ b'_1) \circ \ldots \circ (b_k\ b'_k)$.

By Theorem 10 we can calculate that

$$supp(\pi u) \cap S = (supp(u) \cap A) \cap S \quad \text{and}$$
$$supp(\pi u) \cap A = supp(u) \cap A.$$

Therefore $supp(\pi u) \cap S \subseteq supp(\pi u) \cap A$. Also $(\pi u)\|_A = u\|_A$ by Lemma 13, so $(\pi u, A) \in \mathsf{plane}_S(Z)$. Finally we note that $u \in (\pi u)\|_A$. □

3.3 The substitution action, with examples

We can now define the substitution action. We use Lemma 26 to view an FM set Z as a union of planes; the 'capture-avoiding' aspect of substitution is easy to manage on a 'plane-by-plane basis'.

Definition 27. *If* $A, S \subseteq \mathbb{A}$ *are finite then define*

$$A(a{\mapsto}S) = \begin{cases} (A \setminus \{a\}) \cup S & \text{if } a \in A \\ A & \text{if } a \notin A. \end{cases}$$

Definition 28. *Define the* **substitution action** $z[a{\mapsto}x]$ *and a 'helper' function* $\delta(z, a, x)$ *as follows:*

- $a[a{\mapsto}x] = x$ *and* $b[a{\mapsto}x] = b$, *and*
- *if* $Z \notin \mathbb{A}$ *then*

$$Z[a{\mapsto}x] = \bigcup\{(u[a{\mapsto}x])\|_{A(a\mapsto supp(x))\setminus\delta(u,a,x)} \mid (u, A) \in \mathsf{plane}_{supp(x)\cup\{a\}}(Z)\}$$
$$\delta(u, a, x) = (supp(u)(a{\mapsto}supp(x))) \setminus supp(u[a{\mapsto}x]).$$

We consider some examples.

1. $\{a\}[a{\mapsto}x]$. There is one plane, $(a, \{a\})$.
$$\delta(a, a, x) = \{a\}(a{\mapsto}supp(x)) \setminus supp(x) = \emptyset.$$
$$\{a\}(a{\mapsto}supp(x)) \setminus \delta(a, a, x) = supp(x) \setminus \emptyset = supp(x)$$
$$\{a\}[a{\mapsto}x] = a[a{\mapsto}x]\|_{supp(x)} = x\|_{supp(x)} = x.$$

2. $(\mathbb{A} \setminus \{a\})[a{\mapsto}x]$. One plane is $(b, \{a\})$ where $b\#x$ (the others give the same result).
$$\delta(b, a, x) = \{b\}(a{\mapsto}supp(x)) \setminus \{b\} = \emptyset$$
$$\{a\}(a{\mapsto}supp(x)) \setminus \emptyset = supp(x)$$
$$(\mathbb{A} \setminus \{a\})[a{\mapsto}x] = b\|_{supp(x)} = \mathbb{A} \setminus supp(x)$$

3. $\mathbb{A}[a{\mapsto}x]$. One relevant plane is $(b, \emptyset)$ where $b\#x$ (the others give the same result).
$$\delta(b, a, x) = \emptyset \qquad \emptyset(a{\mapsto}supp(x)) \setminus \emptyset = \emptyset$$
$$\mathbb{A}[a{\mapsto}x] = b\|_{\emptyset} = \mathbb{A}$$

4. $([c]a)[a{\mapsto}x] = \{(b, a), (c, a), (d, a), \ldots\}[a{\mapsto}x]$.
One plane is $((c, a), \{a\})$ where $c\#x$ (if $c \in supp(x)$ then $((c, a), \{a\}) \notin \mathsf{plane}_{supp(x)\cup\{a\}}([c]a)$).
We omit calculations showing that $(c, a)[a{\mapsto}x] = (c, x)$; for a general result see Theorem 32 after these examples.
$$\delta((c, a), a, x) = \{c, a\}(a{\mapsto}supp(x)) \setminus supp((c, x))$$
$$= (supp(x) \cup \{c\}) \setminus (supp(x) \cup \{c\}) = \emptyset$$
$$\{a\}(a{\mapsto}supp(x)) \setminus \emptyset = supp(x)$$
$$([c]a)[a{\mapsto}x] = (c, x)\|_{supp(x)} = [c]x$$
The other planes give the same result.

5. $U_b[a{\mapsto}\{b\}]$ where $U_b = \{a\} \cup \{\{a\}, \{c\}, \{d\}, \ldots\}$ for each b.
Two planes are $a\|_{\{a\}}$ (a plane for $\{a\}$) and $\{a\}\|_{\{b\}}$ (a plane for $\{\{a\}, \{c\}, \{d\}, \ldots\}$).
By calculations similar to the examples above, we calculate that
$$a[a{\mapsto}\{b\}] = \{b\} \quad \text{and}$$
$$\{\{a\}, \{c\}, \{d\}, \ldots\}[a{\mapsto}\{b\}] = \{\{a\}, \{c\}, \{d\}, \ldots\}$$
and that $U_b[a{\mapsto}\{b\}] = U$ where we write
$$U = \{\{a\}, \{b\}, \{c\}, \{d\}, \ldots\}.$$
The other planes give the same results.

6. $([c]U_b)[a{\mapsto}\{b\}] = \{(c, U_b), (d, U_b), \ldots\}[a{\mapsto}\{b\}]$.
One plane is $((c, U_b), \{a, b\})$ (the other planes give the same result). Note that $supp(U) = \emptyset$ and $supp((c, U)) = \{c\}$, so that
$$\delta((c, U_b), a, \{b\}) = \{a, b, c\}(a{\mapsto}\{b\}) \setminus \{c\} = \{b\}$$
$$\{a, b\}(a{\mapsto}\{b\}) \setminus \{b\} = \emptyset$$
$$([c]U_b)[a{\mapsto}\{b\}] = (c, U)\|_{\emptyset} = [c]U.$$

In all other examples δ is equal to $\emptyset$. Here, we see how δ is not equal to $\emptyset$. This corrects for the fact that $supp(U_b[a{\mapsto}\{b\}]) \neq supp(U_b)(a{\mapsto}\{b\})$.

Remark 29. Suppose that $A, S \subseteq \mathbb{A}$ are finite. Note that $A(a{\mapsto}S)$ and $A[a{\mapsto}S]$ do not coincide. For example, $\{a\}(a{\mapsto}\{a\}) = \{a\}$ whereas $\{a\}[a{\mapsto}\{a\}] = \{\{a\}\}$.

Lemma 30. $supp(z[a{\mapsto}x]) \subseteq supp(z)(a{\mapsto}supp(x))$.

Proof. By Theorem 11 $supp(z[a{\mapsto}x]) \subseteq supp(z) \cup \{a\} \cup supp(x)$.

Choose some fresh b (so $b\#z, a, x$). By the axiom (α) $z[a{\mapsto}x] = ((b\ a)z)[b{\mapsto}x]$. By Theorem 11

$$supp(((b\ a)z)[b{\mapsto}x]) \subseteq supp((b\ a)z) \cup \{b\} \cup supp(x).$$

The result follows using Theorem 10. □

Lemma 31. $supp(z[a{\mapsto}x]) \supseteq supp(z)(a{\mapsto}supp(x))$ *need not necessarily hold.*

Proof. A counterexample is $U_b[a{\mapsto}\{b\}]$ above. □

In the terminology of Definition 22, the substitution action is naïve on finite sets:

Theorem 32. *If $Z \notin \mathbb{A}$ and Z is* finite *then* $Z[a{\mapsto}x] = \{z[a{\mapsto}x] \mid z \in Z\}$.

Proof. By definition,

$$Z[a{\mapsto}x] = \bigcup\{(u[a{\mapsto}x])\|_{A(a{\mapsto}supp(x))\setminus\delta(u,a,x)} \mid (u, A) \in \mathsf{plane}_{supp(x)\cup\{a\}}(Z)\}$$

Suppose $(u, A) \in \mathsf{plane}_{supp(x)\cup\{a\}}(Z)$. Since $u\|_A \subseteq Z$ and Z is finite, $u\|_A$ is finite. It follows by part 1 of Lemma 18 that $supp(u) \subseteq A$ and $u\|_A = \{u\}$.

By Lemma 30 $supp(u[a{\mapsto}x]) \subseteq supp(u)(a{\mapsto}supp(x))$, so

$$\delta(u, a, x) = (supp(u)(a{\mapsto}supp(x))) \setminus supp(u[a{\mapsto}x]).$$

It follows by set calculations that

$$supp(u[a{\mapsto}x]) \subseteq A(a{\mapsto}supp(x)) \setminus \delta(u, a, x)$$

and the result follows. □

3.4 The substitution action is a substitution action

We now sketch how substitution satisfies (α), $(\#{\mapsto})$, $(\mathbf{var}{\mapsto})$, $(\mathbf{id}{\mapsto})$, and $(\mathbf{abs}{\mapsto})$, from Definition 21.

Theorem 33 is a useful technical result:

Theorem 33. *$b\#Z$ if and only if for all $(u, A) \in \mathsf{plane}(Z)$ it is the case that $b \notin A$.*

As a corollary $supp(Z) = \bigcup\{A \mid (u, A) \in \mathsf{plane}(Z)\}$.

We may use this result in a slightly different form where we write $b \notin A$ instead of $b\#A$; by part 2 of Lemma 18 these are equivalent.

Proof. By definition if $(u, A) \in \mathsf{plane}(Z)$ then $A \subseteq supp(Z)$. The left-to-right implication follows.

Now suppose that $b\#A$ for every $(u, A) \in \mathsf{plane}(Z)$. Choose any fresh $b'\#Z$. By the first part of this result, $b'\#A$ for every $(u, A) \in \mathsf{plane}(Z)$. Using part 1 of Lemma 26 we reason as follows:

$$\begin{array}{lcl} (b'\ b)Z & = & (b'\ b)\bigcup\{(u\|_A) \mid (u, A) \in \mathsf{plane}(Z)\} \\ & = & \bigcup\{(b'\ b)(u\|_A) \mid (u, A) \in \mathsf{plane}(Z)\} \\ & \overset{\text{Theorem 9}}{=} & \bigcup\{u\|_A \mid (u, A) \in \mathsf{plane}(Z)\} \\ & = & Z \end{array}$$

Now $b \notin supp((b'\ b)Z)$ by Theorem 10 and the fact that $b'\#Z$. The result follows. □

For Theorem 36 we need a technical 'capture-avoidance' result:

Lemma 34. *Suppose that $Z \notin \mathbb{A}$, $a \in \mathbb{A}$, and x is any element.*

Suppose that $B = \{b_1, \ldots, b_n\}$ is a finite set of fresh atoms (so $b_i\#x, Z$ for $1 \le i \le n$). Then

$$Z[a{\mapsto}x] = \bigcup\{(u[a{\mapsto}x])\|_{A(a{\mapsto}supp(x))\setminus\delta(u,a,x)} \mid (u, A) \in \mathsf{plane}_{supp(x)\cup\{a\}\cup B}(Z)\}.$$

Notice the B on the far right subscript.

Proof. A routine calculation demonstrates that if $(u, A) \in \mathsf{plane}(Z)$ and $supp(u)$ 'clashes' with atoms in B, then we can find a $\pi \in \mathsf{fix}(A)$ such that $supp(\pi u)$ does not 'clash' with atoms in B; by Lemma 13 the result follows using Lemma 26. □

Lemma 35. *If (for all b, if $b\#z, x$ then $z[a{\mapsto}x] = ((b\ a)z)[b{\mapsto}x]$), then also (for all b, if $b\#z$ then $z[a{\mapsto}x] = ((b\ a)z)[b{\mapsto}x]$).*

Proof. Choose fresh c (so $c\#z$; also $c\#a, b$ since by our permutative convention c, a, b are distinct atoms). By assumption

$$z[a{\mapsto}x] = ((c\ a)z)[c{\mapsto}x] \quad ((b\ a)z)[b{\mapsto}x] = ((c\ b)(b\ a)z)[c{\mapsto}x].$$

The result follows by Theorem 9. □

Theorem 36 $((\alpha))$**.** $Z[a{\mapsto}x] \subseteq ((b\ a)Z)[b{\mapsto}x]$ *if* $b\#Z$.
As a corollary, for any z if $b\#z$ then $z[a{\mapsto}x] = ((b\ a)z)[b{\mapsto}x]$.

Proof. We first prove the corollary. Suppose $b\#z$; there are two cases depending on whether $z \in \mathbb{A}$:

- Suppose $z \in \mathbb{A}$. Then there are three subcases:
  - (i) $z = a$. $z[a{\mapsto}x] = a[a{\mapsto}x] = x = ((b\ a)a)[b{\mapsto}x] = x$.
  - (ii) $z = b$. This contradicts $b\#z$ so there is nothing to prove.
  - (iii) $z = c$ (where $c \notin \{a, b\}$). $c[a{\mapsto}x] = c = ((b\ a)c)[b{\mapsto}x]$.
- Suppose $Z \notin \mathbb{A}$. By the first part, $Z[a{\mapsto}x] \subseteq ((b\ a)Z)[b{\mapsto}x]$. Also by Theorem 10 $a\#(b\ a)Z$ and it follows that $((b\ a)Z)[b{\mapsto}x] \subseteq ((b\ a)(b\ a)Z)[a{\mapsto}x]$. The result follows, since $(b\ a)(b\ a)Z = Z$.

We now prove by ϵ-induction that $Z[a{\mapsto}x] \subseteq ((b\ a)Z)[b{\mapsto}x]$.

Suppose $Z \notin \mathbb{A}$ and $b\#Z$. By Lemma 35 we can assume $b\#x$. Suppose the inductive hypothesis of every $u \in Z$. We unpack the definition of substitution, using Lemma 34 to add a $\{b\}$ to the subscript on plane in the first equality (we cannot add $\{a\}$ to the subscript on plane in the second equality because we do not know $a\#x$):

$$Z[a{\mapsto}x] = \bigcup\{(u[a{\mapsto}x])\|_{A(a{\mapsto}supp(x))\setminus\delta(u,a,x)} \mid (u, A) \in \mathsf{plane}_{supp(x)\cup\{a,b\}}(Z)\}$$

$$((b\ a)Z)[b{\mapsto}x] = \bigcup\{(u'[b{\mapsto}x])\|_{A'(b{\mapsto}supp(x))\setminus\delta(u',b,x)} \mid (u', A') \in \mathsf{plane}_{supp(x)\cup\{b\}}((b\ a)Z)\}$$

Suppose $(u, A) \in \mathsf{plane}_{supp(x)\cup\{a,b\}}(Z)$. To prove our set inclusion we exhibit $(u', A') \in \mathsf{plane}_{supp(x)\cup\{b\}}((b\ a)Z)$ such that

$$u[a{\mapsto}x]\|_{A(a{\mapsto}supp(x))\setminus\delta(u,a,x)} = u'[b{\mapsto}x]\|_{A'(b{\mapsto}supp(x))\setminus\delta(u',b,x)}.$$

We choose $u' = (b\ a)u$ and $A' = (b\ a)A$. By Theorem 3 we have $(u', A') \in \mathsf{plane}((b\ a)Z)$. Also by definition of $\mathsf{plane}_{supp(x)\cup\{a,b\}}(Z)$ we know that

$$supp(u) \cap (supp(x) \cup \{a, b\}) \subseteq supp(u) \cap A.$$

Now $b \notin A$ by Theorem 33 and $b \in supp(x) \cup \{a, b\}$. Therefore $b\#u$. It is now not hard to use Theorem 10 and some elementary set calculations to calculate that

$$supp(u') \cap (supp(x) \cup \{b\}) \subseteq supp(u') \cap A'.$$

So $(u', A') \in \mathsf{plane}_{supp(x)\cup\{b\}}(Z)$. Also since $b\#u$ by the inductive hypothesis $u'[b{\mapsto}x] = u[a{\mapsto}x]$.

It remains to show

$$A(a{\mapsto}supp(x))\backslash\delta(u,a,x) = ((b\,a)A)(b{\mapsto}supp(x))\backslash\delta((b\,a)u,b,x).$$

Recall that $b \notin A$. Then $A(a{\mapsto}supp(x)) = ((b\,a)A)(b{\mapsto}supp(x))$ is easily verified. Also

$$\delta((b\,a)u,b,x) = supp((b\,a)u)(b{\mapsto}supp(x))\backslash supp(((b\,a)u)[b{\mapsto}x]).$$

Now $supp((b\,a)u)(b{\mapsto}supp(x)) = supp(u)(a{\mapsto}supp(x))$ is easily verified, and $supp(((b\,a)u)[b{\mapsto}x]) = supp(u[a{\mapsto}x])$ follows by the inductive hypothesis. The result follows. □

Theorem 37 $((\#{\mapsto}))$**.** *For all $a \in \mathbb{A}$, if $a\#z$ then $z[a{\mapsto}x] = z$.*

Proof. We work by ϵ-induction. The interesting case is when $Z \notin \mathbb{A}$ (we adhere to our convention and write capital Z) and $a\#Z$. Suppose the inductive hypothesis of all $u \in Z$. By definition

$$Z[a{\mapsto}x] = \bigcup\{(u[a{\mapsto}x])\|_{A(a{\mapsto}supp(x))\backslash\delta(u,a,x)} \mid (u, A) \in \mathsf{plane}_{supp(x)\cup\{a\}}(Z)\}.$$

For any $(u, A) \in \mathsf{plane}_{supp(x)\cup\{a\}}(Z)$ by assumption

$$supp(u) \cap (supp(x) \cup \{a\}) \subseteq supp(u) \cap A.$$

By Theorem 33 $a\#A$, so $a\#u$ and by inductive hypothesis $u[a{\mapsto}x] = u$. Now $A(a{\mapsto}supp(x)) = A$, and

$$\begin{aligned}\delta(u,a,x) &= supp(u)(a{\mapsto}supp(x)) \setminus supp(u[a{\mapsto}x]) \\ &= supp(u) \setminus supp(u) = \emptyset \quad \text{and} \\ Z[a{\mapsto}x] &= \bigcup\{u\|_A \mid (u, A) \in \mathsf{plane}_{supp(x)\cup\{a\}}(Z)\}.\end{aligned}$$

The result follows by part 2 of Lemma 26. □

Theorem 38 $((\mathbf{abs}{\mapsto}))$**.** *If $c\#x$ then $([c]z)[a{\mapsto}x] = [c](z[a{\mapsto}x])$.*

Proof. If $a\#z$ then by Theorem 11 also $a\#[c]z$ and

$$([c]z)[a{\mapsto}x] = [c]z \qquad \text{and} \qquad [c](z[a{\mapsto}x]) = [c]z$$

follow by Theorem 37. So suppose $a \in supp(z)$. We sketch the rest of the proof: It is a fact that

$$((c, z), supp(z) \setminus \{c\}) \in \mathsf{plane}([c]z).$$

The other planes add nothing to the final result. So

$$\begin{aligned}[c]z &= (c, z)\|_{supp(z)\setminus\{c\}} \\ ([c]z)[a{\mapsto}x] &= (c, z[a{\mapsto}x])\|_{(supp(z)\setminus\{c\})(a{\mapsto}supp(x))\backslash\delta((c,z),a,x)} \\ [c](z[a{\mapsto}x]) &= (c, z[a{\mapsto}x])\|_{supp(z[a{\mapsto}x])\setminus\{c\}}.\end{aligned}$$

It suffices to verify that

$$(supp(z)\backslash\{c\})(a{\mapsto}supp(x))\backslash\delta((c,z),a,x) = supp(z[a{\mapsto}x])\backslash\{c\}.$$

Now

$$\delta((c,z),a,x) = (supp(z)\cup\{c\})(a{\mapsto}supp(x))\backslash(supp(z[a{\mapsto}x])\cup\{c\})$$

(we use part 3 of Lemma 18 to calculate the support of a pairset). The result follows by set calculations. □

Theorem 39 $((\mathbf{id}{\mapsto}))$**.** $z[a{\mapsto}a] = z$.

Proof. By an easy inductive argument which we sketch. The interesting case is of $Z \notin \mathbb{A}$ where we suppose $u[a{\mapsto}a] = u$ for all $u \in Z$ (we adhere to our convention and write capital Z). By definition

$$Z[a{\mapsto}a] = \bigcup\{(u[a{\mapsto}a])\|_{A(a{\mapsto}\{a\})\setminus\delta(u,a,a)} \mid (u, A) \in \mathsf{plane}_{\{a\}}(Z)\}.$$

This easily simplifies using the inductive hypothesis to

$$Z[a{\mapsto}a] = \bigcup\{u\|_A \mid (u, A) \in \mathsf{plane}_{\{a\}}(Z)\}$$

and we use Lemma 26. □

Theorem 40. *Definition 28 is equivariant and satisfies (α), $(\#{\mapsto})$, $(\mathbf{var}{\mapsto})$, $(\mathbf{id}{\mapsto})$, and $(\mathbf{abs}{\mapsto})$ from Subsection 3.1.*

Proof. Equivariance is automatic by Theorem 3. $(\mathbf{var}{\mapsto})$ is direct from the definition. Each of (α), $(\#{\mapsto})$, $(\mathbf{id}{\mapsto})$, and $(\mathbf{abs}{\mapsto})$ is by one of the theorems proved above. □

3.5 Substitution and abstract syntax

As a sanity check we prove that our substitution action extends the substitution on syntax, if we express syntax in a model of FM set theory as outlined in previous work [15]. In other words: our substitution action coincides with our expectations of what substitution does to syntax, in a sense which we make precise in Theorem 43.

Definition 41. *Let Λ be inductively defined by:*

$$\frac{a \in \mathbb{A}}{a \in \Lambda} \qquad \frac{x, y \in \Lambda}{(x, y) \in \Lambda} \qquad \frac{a \in \mathbb{A} \quad x \in \Lambda}{[a]x \in \Lambda}$$

Lemma 42. *Λ is isomorphic to λ-terms up to α-equivalence.*

Proof. This is the FM standard construction of abstract-syntax-with-binding [16] slightly modified (see Remark 44 below). □

Theorem 43. *If $z, x \in \Lambda$ then $z[a{\mapsto}x] \in \Lambda$ and $z[a{\mapsto}x]$ is equal to what we usually call 'capture-avoiding substitution of x for a in z'.*

Proof. We work by induction on Λ.

- $a[a{\mapsto}x] = x$ and $b[a{\mapsto}x] = b$.
- $(z_1, z_2)[a{\mapsto}x] = (z_1[a{\mapsto}x], z_2[a{\mapsto}x])$ from Definition 15 and Theorem 32.
- $([c]z)[a{\mapsto}x] = [c](z[a{\mapsto}x])$ providing $c\#x$ by Theorem 38. It is not hard to use part 3 of Lemma 18 and Theorem 19 to prove that $c\#x$ corresponds precisely to 'c is not free in x' when $x \in \Lambda$. □

Define $\mathsf{inl}(x) = (x, 0)$ and $\mathsf{inr}(x) = (x, 1)$.

Remark 44. Theorem 43 works generically for any datatype of syntax-with-binding. Note that we must interpret atoms as themselves and not 'wrapped up': our construction is an isomorphic version of the datatype from [16] given by:

$$\frac{a \in \mathbb{A}}{\mathsf{inl}(\mathsf{inl}(a)) \in \Lambda} \qquad \frac{x, y \in \Lambda}{\mathsf{inl}(\mathsf{inr}((x, y))) \in \Lambda} \qquad \frac{a \in \mathbb{A} \quad x \in \Lambda}{\mathsf{inr}([a]x) \in \Lambda}$$

This is not suitable for Theorem 43 because atoms are wrapped up in $\mathsf{inl}(\mathsf{inl}(a))$ and $\mathsf{inl}(\mathsf{inl}(a))[a{\mapsto}x] = \mathsf{inl}(\mathsf{inl}(x)) \neq x$. There is no canonical implementation of the tree-structure of datatypes — the FM substitution action cannot 'guess' which implementation we chose for inl and inr.

Atoms are a distinct class of elements in a model of FM set theory so it does not harm to insert them 'unwrapped' into Λ.

3.6 Commuting substitutions

It is routine to prove the usual commutativity property of substitutions. The proof is generic and would work for any datatype:

Corollary 45. *If $a\#y$ and $x, y, z \in \Lambda$ then*

$$z[a{\mapsto}x][b{\mapsto}y] = z[b{\mapsto}y][a{\mapsto}x[b{\mapsto}y]].$$

Proof. By induction on z. Only the base case is interesting:

$$a[a{\mapsto}x][b{\mapsto}y] = x[b{\mapsto}y] = a[b{\mapsto}y][a{\mapsto}x[b{\mapsto}y]]. \qquad \square$$

For more complex sets substitutions need not commute. That is:

Lemma 46. *There exist z, a, x, b, y such that $a\#y$ and $z[a{\mapsto}x][b{\mapsto}y] \neq z[b{\mapsto}y][a{\mapsto}x[b{\mapsto}y]]$.*

Proof.
$$\begin{aligned}\{a,\{a\},\{c\},\{d\},\ldots\}[a{\mapsto}\{b\}][b{\mapsto}\{c\}] &= \{\{a\},\{b\},\{c\},\{d\},\ldots\}[b{\mapsto}\{c\}]\\ &= \{\{a\},\{b\},\{c\},\{d\},\ldots\}\end{aligned}$$

$$\begin{aligned}\{a,\{a\},\{c\},\{d\},\ldots\}[b{\mapsto}\{c\}][a{\mapsto}\{\{c\}\}] &= \{a,\{a\},\{b\},\{d\},\ldots\}[a{\mapsto}\{\{c\}\}]\\ &= \{\{a\},\{b\},\{\{c\}\},\{d\},\ldots\}\end{aligned}$$

(The planes of interest here are $a\|_{\{a\}}$, $\{a\}\|_{\{b\}}$, and $\{a\}\|_{\{c\}}$.) $\square$

Intuitively, $\{a,\{a\},\{c\},\{d\},\ldots\}$ can be read as the predicate 'is the variable a, or is $\{x\}$ where x is a variable other than c'; see the Conclusions. Predicates which reflect on their own variables cannot be expressed in standard logics such as first-order logics; non-commutativity of substitution only holds on sets which intuitively 'reflect on atoms' in which case the results obtained may depend on the order in which those atoms are substituted.

4 Conclusions

We have exhibited substitution as an operation in models of FM set theory, with the same status as 'the graph of a function', 'ordered pairs', 'ordinals', and other basic concepts of mathematics. The foundations of computer science are not set in stone, and by paying attention to them, new insights can be gained.

From the philosophical point of view the FM universe provides a basis by which we can obtain a semantics for formal languages where the structure of denotations matches the structure of syntax very closely, also for open terms. In the case that the denotations are *of* formal syntax, the two coincide as exemplified in Section 3.5.

Since our intention here is to the lay the foundations provided by FM set theory, a full semantic treatment of first order logic is beyond the scope of this paper. However we can hint at how it works. An FM model of first order logic maps the sentences $\mathtt{P}, \mathtt{Q}\ldots$ of first order logic, open or closed, into subsets of an FM set U representing the universe of discourse. The model also maps the variables $\mathtt{x}, \mathtt{y}\ldots$ of a first order language into the set $\mathbb{A}$ of atoms. Now, if a first order sentence $\mathtt{P}$ is assigned the set $y \subseteq U$ as its semantic denotation, then the sentence $\forall\mathtt{x}\mathtt{P}$ has as its denotation the set $\bigcap\{y[a{\mapsto}u] \mid u \in U\}$. This is more than just a translation of the syntactic substitution-for-all-terms operation into another language, for $[a{\mapsto}u]$ represents a particular set theoretic operation constructable from the axioms of FM set theory.

Future work is to use the substitution action above as the basis of semantics for formal languages essential to philosophy and computer science — first-order logic and the λ-calculus are two candidates. It is also possible to investigate 'rewriting on sets'; starting with investigating the unifiers of two sets.

REFERENCES

[1] S. Abramsky, D. R. Ghica, A. S. Murawski, C.-H. L. Ong, and I. D. B. Stark, 'Nominal games and full abstraction for the nu-calculus', in *LICS*, pp. 150–159. IEEE, (2004).

[2] Peter Aczel, *Non-wellfounded Set Theory*, number 14 in CSLI lecture notes, CSLI, 1988.

[3] Peter Aczel, 'Generalised set theory', *CSLI lecture notes*, **1**(58), 1–17, (1996).

[4] Peter Aczel and Rachel Lunnon, 'Universes and parameters', *CSLI lecture notes*, **2**, 3–24, (1991).

[5] Nick Benton and Benjamin Leperchey, 'Relational reasoning in a nominal semantics for storage.', in *Proc. of the 7th Int'l Conf. on Typed Lambda Calculi and Applications (TLCA)*, volume 3461 of *LNCS*, pp. 86–101, (2005).

[6] N. Brunner. 75 years of independence proofs by Fraenkel-Mostowski permutation models, 1996.

[7] James Cheney and Christian Urban, 'Alpha-prolog: A logic programming language with names, binding and alpha-equivalence', in *Proc. of the 20th Int'l Conf. on Logic Programming (ICLP 2004)*, eds., Bart Demoen and Vladimir Lifschitz, number 3132 in LNCS, pp. 269–283. Springer-Verlag, (2004).

[8] N. G. de Bruijn, 'Lambda calculus notation with nameless dummies, a tool for automatic formula manipulation, with application to the Church-Rosser theorem', *Indagationes Mathematicae*, **5**(34), 381–392, (1972).

[9] Kit Fine, *Reasoning with Arbitrary Objects*, Blackwell, 1985.

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

[11] Gottlob Frege, 'Begriffsschrift, eine der Arithmetischen Nachgebildete Formelsprache des Reinen Denkens', in *From Frege to Gödel: A Source Book in Mathematical Logic, 1879-1931*, ed., J. van Heijenoort, Harvard University Press, (2002). Translated by S. Bauer-Mengelberg as 'Concept Script, a formal language of pure thought modelled upon that of arithmetic'.

[12] Murdoch J. Gabbay, *A Theory of Inductive Definitions with alpha-Equivalence*, Ph.D. dissertation, Cambridge, UK, 2000.

[13] Murdoch J. Gabbay and Aad Mathijssen, 'Capture-avoiding substitution as a nominal algebra', *Formal Aspects of Computing*, (2008). Available online.

[14] Murdoch J. Gabbay and Aad Mathijssen, 'One-and-a-halfth-order logic', *Journal of Logic and Computation*, (2008). Available online.

[15] Murdoch J. Gabbay and A. M. Pitts, 'A new approach to abstract syntax involving binders', in *14th Annual Symposium on Logic in Computer Science*, pp. 214–224. IEEE Computer Society Press, (1999).

[16] Murdoch J. Gabbay and A. M. Pitts, 'A new approach to abstract syntax with variable binding', *Formal Aspects of Computing*, **13**(3–5), 341–363, (2001).

[17] Thomas Jech, 'Set theory', in *The Stanford Encyclopedia of Philosophy*, ed., Edward N. Zalta, (Fall 2002).

[18] P. T. Johnstone, *Notes on logic and set theory*, Cambridge University Press, 1987.

[19] D. Miller, 'Abstract syntax for variable binders: An overview', *Lecture Notes in Articial Intelligence*, **1861**, 239–253, (July 2000).

[20] A. M. Pitts, 'Nominal logic, a first order theory of names and binding', *Information and Computation*, **186**(2), 165–193, (2003).

[21] A. M. Pitts and Murdoch J. Gabbay, 'A metalanguage for programming with bound names modulo renaming', in *Proceedings of the 5th international conference on the Mathematics of Program Construction (MPC2000)*, eds., R. Backhouse and J. N. Oliveira, volume 1837 of *LNCS*, pp. 230–255. Springer, (July 2000).

[22] J. Truss, 'Permutations and the axiom of choice', in *Automorphisms of first order structures*, ed., H.D.Macpherson R.Kaye, 131–152, OUP, (1994).

[23] Christian Urban and Christine Tasson, 'Nominal techniques in Isabelle/HOL', in *CADE 2005*, volume 3632 of *Lecture Notes in Artificial Intelligence*, pp. 38–53, (2005).