AISB/IACAP World Congress 2012

Birmingham, UK, 2-6 July 2012

Computational Philosophy

Anthony F. Beavers (Editor)

Part of

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

http://www.aisb.org.uk

ISBN 978-1-908187-12-3

Foreword from the Congress Chairs

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

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

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

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

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

Contents

Evidence-Based Interpretations of PA

Bhupinder Singh Anand

Abstract. We show that Tarski's inductive definitions admit evidence-based interpretations of the first-order Peano Arithmetic PA that allow us to define the satisfaction and truth of the quantified formulas of PA *constructively* over the domain N of the natural numbers in *two* essentially different ways: (1) in terms of algorithmic verifiabilty; and (2) in terms of algorithmic computability. We argue that the algorithmically computable PA-formulas *can* provide a finitary interpretation of PA over the domain N of the natural numbers from which we may conclude that PA is consistent.

Keywords Algorithmic computability, algorithmic verifiability, Aristotle's particularisation, atomic formulas, consistency, constructive, decidability, domain, finitary, finite induction, first-order, formal language, interpretation, natural numbers, numerals, Gödel, ω-consistency, Peano Arithmetic PA, satisfaction, soundness, standard interpretation, Tarski, truth.

1 Introduction

In this paper we seek to address one of the philosophical challenges associated with accepting arithmetical propositions as true under an interpretation—either axiomatically or on the basis of subjective self-evidence—without any effective methodology for objectively evidencing such acceptance[1].

For instance, conventional wisdom accepts Alfred Tarski's definitions of the satisfiability and truth of the formulas of a formal language under an interpretation[2] and *postulates* that, under the standard interpretation $\mathcal{I}_{PA(N,\ Standard)}$ of the first-order Peano Arithmetic PA[3] over the domain N of the natural numbers:

(i) The atomic formulas of PA can be *assumed* as decidable under $\mathcal{I}_{PA(N,\ Standard)}$;

(ii) The PA axioms can be *assumed* to interpret as satisfied/true under $\mathcal{I}_{PA(N,\ Standard)}$;

(iii) the PA rules of inference—Generalisation and Modus Ponens—can be *assumed* to preserve such satisfaction/truth under $\mathcal{I}_{PA(N,\ Standard)}$.

> **Standard interpretation of PA** The standard interpretation $\mathcal{I}_{PA(N,\ Standard)}$ of PA over the domain N of the natural numbers is the one in which the logical constants have their 'usual' interpretations[4] in Aristotle's logic of predicates (which subsumes Aristotle's particularisation[5]), and[6]:
>
> (a) the set of non-negative integers is the domain;
> (b) the symbol [0] interprets as the integer 0;
> (c) the symbol [$'$] interprets as the successor operation (addition of 1);
> (d) the symbols [+] and [$\star$] interpret as ordinary addition and multiplication;
> (e) the symbol [=] interprets as the identity relation.
>
> **The axioms of first-order Peano Arithmetic (PA)**
>
> **PA**$_1$ $[(x_1 = x_2) \rightarrow ((x_1 = x_3) \rightarrow (x_2 = x_3))]$;
> **PA**$_2$ $[(x_1 = x_2) \rightarrow (x_1' = x_2')]$;
> **PA**$_3$ $[0 \neq x_1']$;
> **PA**$_4$ $[(x_1' = x_2') \rightarrow (x_1 = x_2)]$;
> **PA**$_5$ $[(x_1 + 0) = x_1]$;
> **PA**$_6$ $[(x_1 + x_2') = (x_1 + x_2)']$;
> **PA**$_7$ $[(x_1 \star 0) = 0]$;
> **PA**$_8$ $[(x_1 \star x_2') = ((x_1 \star x_2) + x_1)]$;
> **PA**$_9$ For any well-formed formula $[F(x)]$ of PA:
> $[F(0) \rightarrow (((\forall x)(F(x) \rightarrow F(x'))) \rightarrow (\forall x)F(x))]$.
>
> **Generalisation in PA** If $[A]$ is PA-provable, then so is $[(\forall x)A]$.
>
> **Modus Ponens in PA** If $[A]$ and $[A \rightarrow B]$ are PA-provable, then so is $[B]$.

We shall show that although the seemingly innocent and self-evident assumption in (i) can, indeed, be justified, it conceals an ambiguity whose impact on (ii) and (iii) is far-reaching in significance and needs to be made explicit.

Reason: Tarski's inductive definitions admit evidence-based interpretations of PA that actually allow us to metamathematically define the satisfaction and truth of the atomic (and, ipso facto, quantified) formulas of PA *constructively* over N in *two* essentially different ways as below, only one of which is *finitary*[7]:

(1) in terms of algorithmic *verifiabilty*[8];

(2) in terms of algorithmic *computability*[9].

---

[1] For a brief recent review of such challenges, see [Fe06], [Fe08].
[2] As detailed in Section 4.
[3] We take this to be the first-order theory S defined in [Me64], p.102.
[4] We essentially follow the definitions in [Me64], p.49.
[5] We define this important concept explicitly later in Section 2.1. Loosely speaking, Aristotle's particularisation is the assumption that we may always interpret the formal expression '$[(\exists x)F(x)]$' of a formal language under an interpretation as 'There exists an object s in the domain of the interpretation such that $F(s)$'.
[6] See [Me64], p.107.
[7] 'Finitary' in the sense that "...there should be an algorithm for deciding the truth or falsity of any mathematical statement" ... http://en.wikipedia.org/wiki/Hilbert's_program. For a brief review of 'finitism' and 'constructivity' in the context of this paper see [Fe08].
[8] Section 3, Definition 1.
[9] Section 3, Definition 2.

Case 1: We show in Section 4.2 that the algorithmically verifiable PA-formulas admit an unusual, 'instantiational' Tarskian interpretation $\mathcal{I}_{PA(\mathbb{N},\ Instantiational)}$ of PA over the domain $\mathbb{N}$ of the PA *numerals*; and that this interpretation is sound if, and only if, PA is ω-consistent.

> **Soundness (formal system):** We define a formal system S as sound under a Tarskian interpretation $\mathcal{I}_S$ over a domain D if, and only if, every theorem $[T]$ of S translates as '$[T]$ is true under $\mathcal{I}_S$ in D'.
>
> **Soundness (interpretation):** We define a Tarskian interpretation $\mathcal{I}_S$ of a formal system S as sound over a domain D if, and only if, S is sound under the interpretation $\mathcal{I}_S$ over the domain D.
>
> **Simple consistency:** A formal system S is simply consistent if, and only if, there is no S-formula $[F(x)]$ for which both $[(\forall x)F(x)]$ and $[\neg(\forall x)F(x)]$ are S-provable.
>
> **ω-consistency:** A formal system S is ω-consistent if, and only if, there is no S-formula $[F(x)]$ for which, first, $[\neg(\forall x)F(x)]$ is S-provable and, second, $[F(a)]$ is S-provable for any given S-term $[a]$.

We further show that this interpretation can be viewed as a formalisation of the standard interpretation $\mathcal{I}_{PA(N,\ Standard)}$ of PA over N; in the sense that—under Tarski's definitions—$\mathcal{I}_{PA(\mathbb{N},\ Instantiational)}$ is sound over $\mathbb{N}$ if, and only if, $\mathcal{I}_{PA(N,\ Standard)}$ is sound over N (as postulated in (ii) and (iii) above).

Although the standard interpretation $\mathcal{I}_{PA(N,\ Standard)}$ is *assumed* to be sound over N (as expressed by (ii) and (iii) above), it cannot claim to be finitary since it it is not known to lead to a finitary justification of the truth—under Tarski's definitions—of the Axiom Schema of (finite) Induction of PA in N from which we may conclude—in an intuitionistically unobjectionable manner—that PA is consistent[10].

> We note that Gerhard Gentzen's 'constructive'[11] consistency proof for formal number theory[12] is debatably finitary[13], since it involves a Rule of Infinite Induction that appeals to the properties of transfinite ordinals.

Case 2: We show further in Section 4.3 that the algorithmically computable PA-formulas admit an 'algorithmic' Tarskian interpretation $\mathcal{I}_{PA(N,\ Algorithmic)}$ of PA over N.

We then argue in Section 5 that $\mathcal{I}_{PA(N,\ Algorithmic)}$ is essentially different from $\mathcal{I}_{PA(\mathbb{N},\ Instantiational)}$ since the PA-axioms—including the Axiom Schema of (finite) Induction—are algorithmically computable as satisfied/true under the standard interpretation of PA over N, and the PA rules of inference preserve algorithmically computable satisfiability/truth under the interpretation[14].

We conclude from the above that the interpretation $\mathcal{I}_{PA(N,\ Algorithmic)}$ is finitary, and hence sound over N[15].

We further conclude from the soundness of the interpretation $\mathcal{I}_{PA(N,\ Algorithmic)}$ over N that PA is consistent[16].

[10] The possibility/impossibility of such justification was the subject of the famous Poincaré-Hilbert debate. See [Hi27], p.472; also [Br13], p.59; [We27], p.482; [Pa71], p.502-503.
[11] In the sense highlighted by Elliott Mendelson in [Me64], p.261.
[12] cf. [Me64], p258.
[13] See for instance http://en.wikipedia.org/wiki/Hilbert's_program.
[14] Section 5.2, Theorem 4.
[15] Section 5.3, Theorem 5.
[16] Section 5.3, Theorem 6.

2 Interpretation of an arithmetical language in terms of the computations of a simple functional language

We begin by noting that we can, in principle, define[17] the classical 'satisfaction' and 'truth' of the formulas of a first order arithmetical language, such as PA, *verifiably* under an interpretation using as *evidence*[18] the computations of a simple functional language.

Such definitions follow straightforwardly for the atomic formulas of the language (i.e., those without the logical constants that correspond to 'negation', 'conjunction', 'implication' and 'quantification') from the standard definition of a simple functional language[19].

Moreover, it follows from Alfred Tarski's seminal 1933 paper on the concept of truth in the languages of the deductive sciences[20] that the 'satisfaction' and 'truth' of those formulas of a first-order language which contain logical constants can be inductively defined, under an interpretation, in terms of the 'satisfaction' and 'truth' of the interpretations of only the atomic formulas of the language.

Hence the 'satisfaction' and 'truth' of those formulas (of an arithmetical language) which contain logical constants can, in principle, also be defined verifiably under an interpretation using as evidence the computations of a simple functional language.

We show in Section 4 that this is indeed the case for PA under its standard interpretation $\mathcal{I}_{PA(N,\ Standard)}$, when this is explicitly defined as in Section 5.

We show, moreover, that we can further define 'algorithmic truth' and 'algorithmic falsehood' under $\mathcal{I}_{PA(N,\ Standard)}$ such that the PA axioms interpret as always algorithmically true, and the rules of inference preserve algorithmic truth, over the domain N of the natural numbers.

2.1 The definitions of 'algorithmic truth' and 'algorithmic falsehood' under $\mathcal{I}_{PA(N,\ Standard)}$ are not symmetric with respect to 'truth' and 'falsehood' under $\mathcal{I}_{PA(N,\ Standard)}$

However, the definitions of 'algorithmic truth' and 'algorithmic falsehood' under $\mathcal{I}_{PA(N,\ Standard)}$ are not symmetric with respect to classical (verifiable) 'truth' and 'falsehood' under $\mathcal{I}_{PA(N,\ Standard)}$.

For instance, if a formula $[(\forall x)F(x)]$ of an arithmetic is algorithmically true under an interpretation (such as $\mathcal{I}_{PA(N,\ Standard)}$), then we may conclude that there is an algorithm that, for any given numeral $[a]$, provides evidence that the formula $[F(a)]$ is algorithmically true under the interpretation.

In other words, there is an algorithm that provides evidence that the interpretation $F^*(a)$ of $[F(a)]$ *holds* in N for any given natural number a.

> **Notation:** We use enclosing square brackets as in '$[F(x)]$' to indicate that the expression inside the brackets is to be

[17] Formal definitions are given in Section 4.
[18] [Mu91].
[19] Such as, for instance, that of a deterministic Turing machine ([Me64], pp.229-231) based essentially on Alan Turing's seminal 1936 paper on computable numbers ([Tu36]).
[20] [Ta33].

treated as denoting a formal expression (formal string) of a formal language. We use an asterisk as in '$F^*(x)$' to indicate the asterisked expression $F^*(x)$ is to be treated as denoting the interpretation of the formula $[F(x)]$ in the corresponding domain of the interpretation.

Defining the term 'hold': We define the term 'hold'—when used in connection with an interpretation of a formal language L and, more specifically, with reference to the computations of a simple functional language associated with the atomic formulas of the language L—explicitly in Section 4; the aim being to avoid appealing to the classically subjective (and existential) connotation implicitly associated with the term under an implicitly defined standard interpretation of an arithmetic[21].

However, if a formula $[(\forall x)F(x)]$ of an arithmetic is algorithmically false under an interpretation, then we can only conclude that there is no algorithm that, for any given natural number a, can provide evidence whether the interpretation $F^*(a)$ holds or not in N .

We cannot conclude that there is a numeral $[a]$ such that the formula $[F(a)]$ is algorithmically false under the interpretation; nor can we conclude that there is a natural number b such that $F^*(b)$ does not hold in N.

Such a conclusion would require:

(i) either some additional evidence that will verify for some assignment of numerical values to the free variables of $[F]$ that the corresponding interpretation F^* does not hold[22];

(ii) or the additional assumption that either Aristotle's particularisation holds over the domain of the interpretation (as is implicitly presumed under the standard interpretation of PA over N) or, equivalently, that the arithmetic is ω-consistent[23].

Aristotle's particularisation This holds that from a meta-assertion such as:

'It is not the case that: For any given x, $P^*(x)$ does not hold',

usually denoted symbolically by '$\neg(\forall x)\neg P^*(x)$', we may always validly infer in the classical, Aristotlean, logic of predicates[24] that:

'There exists an unspecified x such that $P^*(x)$ holds',

usually denoted symbolically by '$(\exists x)P^*(x)$'.

The significance of Aristotle's particularisation for the first-order predicate calculus: We note that in a formal language the formula '$[(\exists x)P(x)]$' is an abbreviation for the formula '$[\neg(\forall x)\neg P(x)]$'. The commonly accepted interpretation of this formula—and a fundamental tenet of classical logic unrestrictedly adopted as intuitively obvious by standard literature[25] that seeks to build upon the formal first-order predicate calculus—tacitly appeals to Aristotlean particularisation.

However, L. E. J. Brouwer had noted in his seminal 1908 paper on the unreliability of logical principles[26] that the commonly accepted interpretation of this formula is ambiguous if interpretation is intended over an infinite domain.

Brouwer essentially argued that, even supposing the formula '$[P(x)]$' of a formal Arithmetical language interprets as an arithmetical relation denoted by '$P^*(x)$', and the formula '$[\neg(\forall x)\neg P(x)]$' as the arithmetical proposition denoted by '$\neg(\forall x)\neg P^*(x)$', the formula '$[(\exists x)P(x)]$' need not interpret as the arithmetical proposition denoted by the usual abbreviation '$(\exists x)P^*(x)$'; and that such postulation is invalid as a general logical principle in the absence of a means for constructing some putative object a for which the proposition $P^*(a)$ holds in the domain of the interpretation.

Hence we shall follow the convention that the assumption that '$(\exists x)P^*(x)$' is the intended interpretation of the formula '$[(\exists x)P(x)]$'—which is essentially the assumption that Aristotle's particularisation holds over the domain of the interpretation—must always be explicit.

The significance of Aristotle's particularisation for PA: In order to avoid intuitionistic objections to his reasoning, Kurt Gödel introduced the syntactic property of ω-consistency as an explicit assumption in his formal reasoning in his seminal 1931 paper on formally undecidable arithmetical propositions[27].

Gödel explained at some length[28] that his reasons for introducing ω-consistency explicitly was to avoid appealing to the semantic concept of classical arithmetical truth in Aristotle's logic of predicates (which presumes Aristotle's particularisation).

It is straightforward to show that the two concepts are metamathematically equivalent in the sense that, if PA is consistent, then PA is ω-consistent if, and only if, Aristotle's particularisation holds under the standard interpretation of PA over N.

3 Defining algorithmic verifiability and algorithmic computability

The asymmetry of Section 2.1 suggests the following two concepts[29]:

Definition 1 ***Algorithmic verifiability:***

An arithmetical formula $[(\forall x)F(x)]$ is algorithmically verifiable as true under an interpretation if, and only if, for any given numeral $[a]$, we can define an algorithm which provides evidence that $[F(a)]$ interprets as true under the interpretation.

Tarskian interpretation of an arithmetical language *verifiably* in terms of the computations of a simple functional language We show in Section 4 that the 'algorithmic verifiability' of the formulas of a formal language which contain logical constants can be inductively defined under an interpretation in terms of the 'algorithmic verifiability' of the interpretations of the atomic formulas of the language; further, that the PA-formulas are decidable under the standard interpretation of PA over N if, and only if, they are algorithmically verifiable under the interpretation (Corollary 2).

Definition 2 ***Algorithmic computability:***

An arithmetical formula $[(\forall x)F(x)]$ is algorithmically computable as true under an interpretation if, and only if, we can define an algorithm that, for any given numeral $[a]$, provides evidence that $[F(a)]$ interprets as true under the interpretation.

---

[21] As, for instance, in [Go31].

[22] Essentially reflecting Brouwer's objection to the assumption of Aristotle's particularisation over an infinite domain.

[23] An assumption explicitly introduced by Gödel in [Go31].

[24] [HA28], pp.58-59.

[25] See [Hi25], p.382; [HA28], p.48; [Sk28], p.515; [Go31], p.32.; [Kl52], p.169; [Ro53], p.90; [BF58], p.46; [Be59], pp.178 & 218; [Su60], p.3; [Wa63], p.314-315; [Qu63], pp.12-13; [Kn63], p.60; [Co66], p.4; [Me64], p.52(ii); [Nv64], p.92; [Li64], p.33; [Sh67], p.13; [Da82], p.xxv; [Rg87], p.xvii; [EC89], p.174; [Mu91]; [Sm92], p.18, Ex.3; [BBJ03], p.102; [Cr05], p.6.

[26] [Br08].

[27] [Go31], p.23 and p.28.

[28] In his introduction on p.9 of [Go31].

[29] The distinction sought to be made between algorithmic verifiabilty and algorithmic computability can be viewed as reflecting in number theory the similar distinction in analysis between, for instance, continuous functions ([Ru53], p.65, §4.5) and uniformly continuous functions ([Ru53], p.65, §4.13); or that between convergent sequences ([Ru53], p.65, §7.1) and uniformly convergent sequences ([Ru53], p.65, §7.7).

Tarskian interpretation of an arithmetical language *algorithmically* in terms of the computations of a simple functional language We show in Section 4 that the 'algorithmic computability' of the formulas of a formal language which contain logical constants can also be inductively defined under an interpretation in terms of the 'algorithmic computability' of the interpretations of the atomic formulas of the language; further, that the PA-formulas are decidable under an algorithmic interpretation of PA over N if, and only if, they are algorithmically computable under the interpretation .

We now show that the above concepts are well-defined under the standard interpretation of PA over N.

4 The *implicit* Satisfaction condition in Tarski's inductive assignment of truth-values under an interpretation

We first consider the significance of the *implicit* Satisfaction condition in Tarski's inductive assignment of truth-values under an interpretation.

We note that—essentially following standard expositions[30] of Tarski's inductive definitions on the 'satisfiability' and 'truth' of the formulas of a formal language under an interpretation—we can define:

Definition 3 *If* $[A]$ *is an atomic formula* $[A(x_1, x_2, \ldots, x_n)]$ *of a formal language S, then the denumerable sequence* $(a_1, a_2, \ldots)$ *in the domain D of an interpretation* $\mathcal{I}_{S(D)}$ *of S satisfies* $[A]$ *if, and only if:*

(i) $[A(x_1, x_2, \ldots, x_n)]$ *interprets under* $\mathcal{I}_{S(D)}$ *as a unique relation* $A^*(x_1, x_2, \ldots, x_n)$ *in D for any witness* $\mathcal{W}_D$ *of D;*

(ii) there is a Satisfaction Method, SM($\mathcal{I}_{S(D)}$*) that provides objective evidence*[31] *by which any witness* $\mathcal{W}_D$ *of D can objectively* ***define*** *for any atomic formula* $[A(x_1, x_2, \ldots, x_n)]$ *of S, and any given denumerable sequence* $(b_1, b_2, \ldots)$ *of D, whether the proposition* $A^*(b_1, b_2, \ldots, b_n)$ *holds or not in D;*

(iii) $A^*(a_1, a_2, \ldots, a_n)$ *holds in D for any* $\mathcal{W}_D$*.*

Witness: From a constructive perspective, the existence of a 'witness' as in (i) above is implicit in the usual expositions of Tarski's definitions.

Satisfaction Method: From a constructive perspective, the existence of a Satisfaction Method as in (ii) above is also implicit in the usual expositions of Tarski's definitions.

A constructive perspective: We highlight the word '***define***' in (ii) above to emphasise the constructive perspective underlying this paper; which is that the concepts of 'satisfaction' and 'truth' under an interpretation are to be explicitly viewed as objective assignments by a convention that is witness-independent. A Platonist perspective would substitute 'decide' for 'define', thus implicitly suggesting that these concepts can 'exist', in the sense of needing to be discovered by some witness-dependent means—eerily akin to a 'revelation'—if the domain D is N.

We can now inductively assign truth values of 'satisfaction', 'truth', and 'falsity' to the compound formulas of a first-order theory S under the interpretation $\mathcal{I}_{S(D)}$ in terms of *only* the satisfiability of the atomic formulas of S over D as usual[32]:

Definition 4 *A denumerable sequence s of D satisfies* $[\neg A]$ *under* $\mathcal{I}_{S(D)}$ *if, and only if, s does not satisfy* $[A]$*;*

Definition 5 *A denumerable sequence s of D satisfies* $[A \rightarrow B]$ *under* $\mathcal{I}_{S(D)}$ *if, and only if, either it is not the case that s satisfies* $[A]$*, or s satisfies* $[B]$*;*

Definition 6 *A denumerable sequence s of D satisfies* $[(\forall x_i)A]$ *under* $\mathcal{I}_{S(D)}$ *if, and only if, given any denumerable sequence t of D which differs from s in at most the i'th component, t satisfies* $[A]$*;*

Definition 7 *A well-formed formula* $[A]$ *of D is true under* $\mathcal{I}_{S(D)}$ *if, and only if, given any denumerable sequence t of D, t satisfies* $[A]$*;*

Definition 8 *A well-formed formula* $[A]$ *of D is false under* $\mathcal{I}_{S(D)}$ *if, and only if, it is not the case that* $[A]$ *is true under* $\mathcal{I}_{S(D)}$*.*

It follows that[33]:

Theorem 1 *(Satisfaction Theorem) If, for any interpretation* $\mathcal{I}_{S(D)}$ *of a first-order theory S, there is a Satisfaction Method SM(*$\mathcal{I}_{S(D)}$*) which holds for a witness* $\mathcal{W}_D$ *of D, then:*

(i) The Δ_0 *formulas of S are decidable as either true or false over D under* $\mathcal{I}_{S(D)}$*;*

(ii) If the Δ_n *formulas of S are decidable as either true or as false over D under* $\mathcal{I}_{S(D)}$*, then so are the* $\Delta(n+1)$ *formulas of S.*

Proof It follows from the above definitions that:

(a) If, for any given atomic formula $[A(x_1, x_2, \ldots, x_n)]$ of S, it is decidable by $\mathcal{W}_D$ whether or not a given denumerable sequence $(a_1, a_2, \ldots)$ of D satisfies $[A(x_1, x_2, \ldots, x_n)]$ in D under $\mathcal{I}_{S(D)}$ then, for any given compound formula $[A^1(x_1, x_2, \ldots, x_n)]$ of S containing any one of the logical constants $\neg, \rightarrow, \forall$, it is decidable by $\mathcal{W}_D$ whether or not $(a_1, a_2, \ldots)$ satisfies $[A^1(x_1, x_2, \ldots, x_n)]$ in D under $\mathcal{I}_{S(D)}$;

(b) If, for any given compound formula $[B^n(x_1, x_2, \ldots, x_n)]$ of S containing n of the logical constants $\neg, \rightarrow, \forall$, it is decidable by $\mathcal{W}_D$ whether or not a given denumerable sequence $(a_1, a_2, \ldots)$ of D satisfies $[B^n(x_1, x_2, \ldots, x_n)]$ in D under $\mathcal{I}_{S(D)}$ then, for any given compound formula $[B^{(n+1)}(x_1, x_2, \ldots, x_n)]$ of S containing $n+1$ of the logical constants $\neg, \rightarrow, \forall$, it is decidable by $\mathcal{W}_D$ whether or not $(a_1, a_2, \ldots)$ satisfies $[B^{(n+1)}(x_1, x_2, \ldots, x_n)]$ in D under $\mathcal{I}_{S(D)}$;

We thus have that:

(c) The Δ_0 formulas of S are decidable by $\mathcal{W}_D$ as either true or false over D under $\mathcal{I}_{S(D)}$;

(d) If the Δ_n formulas of S are decidable by $\mathcal{W}_D$ as either true or as false over D under $\mathcal{I}_{S(D)}$, then so are the $\Delta(n+1)$ formulas of S. □

In other words, if the atomic formulas of of S interpret under $\mathcal{I}_{S(D)}$ as decidable with respect to the Satisfaction Method SM($\mathcal{I}_{S(D)}$) by a witness $\mathcal{W}_D$ over some domain D, then the

[30] cf. [Me64], p.51.
[31] In the sense of [Mu91].
[32] See [Me64], p.51; [Mu91].
[33] cf. [Me64], pp.51-53.

propositions of S (i.e., the Π_n and Σ_n formulas of S) also interpret as decidable with respect to $\mathrm{SM}(\mathcal{I}_{S(D)})$ by the witness $\mathcal{W}_D$ over D.

We now consider the application of Tarski's definitions to various interpretations of first-order Peano Arithmetic PA.

4.1 The standard interpretation of PA over the domain N of the natural numbers

The standard interpretation $\mathcal{I}_{PA(N,\ Standard)}$ of PA over the domain N of the natural numbers is obtained if, in $\mathcal{I}_{S(D)}$:

(a) we define S as PA with standard first-order predicate calculus as the underlying logic[34];
(b) we define D as the set N of natural numbers;
(c) for any atomic formula $[A(x_1, x_2, \ldots, x_n)]$ of PA and sequence $(a_1, a_2, \ldots, a_n)$ of N, we take $\|\mathrm{SATCON}(\mathcal{I}_{PA(N)})\|$ as:

> $\|A^*(a_1^*, a_2^*, \ldots, a_n^*)$ holds in N and, for any given sequence $(b_1^*, b_2^*, \ldots, b_n^*)$ of N, the proposition $A^*(b_1^*, b_2^*, \ldots, b_n^*)$ is decidable in $N\|$;

(d) we define the witness $\mathcal{W}_{(N,\ Standard)}$ informally as the 'mathematical intuition' of a human intelligence for whom, classically, $\|\mathrm{SATCON}(\mathcal{I}_{PA(N)})\|$ has been *implicitly* accepted as *objectively* 'decidable' in N;

> We shall show that such acceptance is justified, but needs to be made explicit since:
>
> **Lemma 1** *$A^*(x_1, x_2, \ldots, x_n)$ is both algorithmically verifiable and algorithmically computable in N by $\mathcal{W}_{(N,\ Standard)}$.*
>
> **Proof** (i) It follows from the argument in Theorem 2 (below) that $A^*(x_1, x_2, \ldots, x_n)$ is algorithmically verifiable in N by $\mathcal{W}_{(N,\ Standard)}$.
>
> (ii) It follows from the argument in Theorem 3 (below) that $A^*(x_1, x_2, \ldots, x_n)$ is algorithmically computable in N by $\mathcal{W}_{(N,\ Standard)}$. The lemma follows. □
>
> Now, although it is not immediately obvious from the standard interpretation of PA over N which of (i) or (ii) may be taken for *explicitly* deciding $\|\mathrm{SATCON}(\mathcal{I}_{PA(N)})\|$ by the witness $\mathcal{W}_{(N,\ Standard)}$, we shall show in Section 4.2 that (i) is consistent with (e) below; and in Section 4.3 that (ii) is inconsistent with (e). Thus the standard interpretation of PA over N implicitly presumes (i).

(e) we postulate that Aristotle's particularisation holds over N[35].

Clearly, (e) does not form any part of Tarski's inductive definitions of the satisfaction, and truth, of the formulas of PA under the above interpretation. Moreover, its inclusion makes $\mathcal{I}_{PA(N,\ Standard)}$ extraneously non-finitary[36].

We note further that if PA is ω-*in*consistent, then Aristotle's particularisation does not hold over N, and the interpretation $\mathcal{I}_{PA(N,\ Standard)}$ is not sound over N.

[34] Where the string $[(\exists \ldots)]$ is defined as—and is to be treated as an abbreviation for—the string $[\neg(\forall \ldots)\neg]$. We do not consider the case where the underlying logic is Hilbert's formalisation of Aristotle's logic of predicates in terms of his ϵ-operator ([Hi27], pp.465-466).

[35] Hence a PA formula such as $[(\exists x)F(x)]$ interprets under $\mathcal{I}_{PA(N,\ Standard)}$ as 'There is some natural number n such that $F(n)$ holds in N.

[36] [Br08].

4.2 An instantiational interpretation of PA over the domain $\mathbb{N}$ of the PA numerals

We next consider an instantiational interpretation $\mathcal{I}_{PA(\mathbb{N},\ Instantiational)}$ of PA over the domain $\mathbb{N}$ of the PA numerals[37] where:

(a) we define S as PA with standard first-order predicate calculus as the underlying logic;
(b) we define D as the set $\mathbb{N}$ of PA numerals;
(c) for any atomic formula $[A(x_1, x_2, \ldots, x_n)]$ of PA and any sequence $[(a_1, a_2, \ldots, a_n)]$ of PA numerals in $\mathbb{N}$, we take $\|\mathrm{SATCON}(\mathcal{I}_{PA(\mathbb{N})})\|$ as:

> $\|[A(a_1, a_2, \ldots, a_n)]$ is provable in PA and, for any given sequence of numerals $[(b_1, b_2, \ldots, b_n)]$ of PA, the formula $[A(b_1, b_2, \ldots, b_n)]$ is decidable as either provable or not provable in PA$\|$;

(d) we define the witness $\mathcal{W}_{(\mathbb{N},\ Instantiational)}$ as the meta-theory $\mathcal{M}_{PA}$ of PA.

> **Lemma 2** *$[A(x_1, x_2, \ldots, x_n)]$ is always algorithmically verifiable in PA by $\mathcal{W}_{(\mathbb{N},\ Instantiational)}$.*
>
> **Proof** It follows from Gödel's definition of the primitive recursive relation xBy[38]—where x is the Gödel number of a proof sequence in PA whose last term is the PA formula with Gödel-number y—that, if $[A(x_1, x_2, \ldots, x_n)]$ is an atomic formula of PA, $\mathcal{M}_{PA}$ can algorithmically verify for any given sequence $[(b_1, b_2, \ldots, b_n)]$ of PA numerals which one of the PA formulas $[A(b_1, b_2, \ldots, b_n)]$ and $[\neg A(b_1, b_2, \ldots, b_n)]$ is necessarily PA-provable. □

Now, if PA is consistent but not ω-consistent, then there is a Gödelian formula $[R(x)]$[39] such that:

(i) $[(\forall x)R(x)]$ is not PA-provable;
(ii) $[\neg(\forall x)R(x)]$ is PA-provable;
(iii) for any given numeral $[n]$, the formula $[R(n)]$ is PA-provable.

However, if $\mathcal{I}_{PA(\mathbb{N},\ Instantiational)}$ is sound over $\mathbb{N}$, then (ii) implies contradictorily that it is not the case that, for any given numeral $[n]$, the formula $[R(n)]$ is PA-provable.

It follows that if $\mathcal{I}_{PA(\mathbb{N},\ Instantiational)}$ is sound over $\mathbb{N}$, then PA is ω-consistent and, ipso facto, Aristotle's particularisation must hold over N.

Moreover, if PA is consistent, then every PA-provable formula interprets as true under some sound interpretation of PA over N. Hence $\mathcal{M}_{PA}$ can effectively decide whether, for any given sequence of natural numbers $(b_1^*, b_2^*, \ldots, b_n^*)$ in N, the proposition $A^*(b_1^*, b_2^*, \ldots, b_n^*)$ holds or not in N.

It follows that $\mathcal{I}_{PA(\mathbb{N},\ Instantiational)}$ can be viewed as a constructive formalisation of the 'standard' interpretation $\mathcal{I}_{PA(N,\ Standard)}$ of PA in which we do not need to non-constructively assume that Aristotle's particularisation holds over N.

[37] The raison d'être, and significance, of such interpretation is outlined in this short unpublished note accessible at http://alixcomsi.com/8_Meeting_Wittgenstein_requirement_1000.pdf.

[38] [Go31], p. 22(45).

[39] Gödel constructively defines, and refers to, this formula by its Gödel number 'r': see [Go31], p.25, Eqn.(12);.

4.3 An algorithmic interpretation of PA over the domain N of the natural numbers

We finally consider the purely algorithmic interpretation $\mathcal{I}_{PA(N,\ Algorithmic)}$ of PA over the domain N of the natural numbers where:

(a) we define S as PA with standard first-order predicate calculus as the underlying logic;
(b) we define D as the set N of natural numbers;
(c) for any atomic formula $[A(x_1, x_2, \ldots, x_n)]$ of PA and any sequence $(a_1, a_2, \ldots, a_n)$ of natural numbers in N, we take $\|\text{SATCON}(\mathcal{I}_{PA(N)})\|$ as:

$\|A^*(a_1^*, a_2^*, \ldots, a_n^*)$ holds in N and, for any given sequence $(b_1^*, b_2^*, \ldots, b_n^*)$ of N, the proposition $A^*(b_1^*, b_2^*, \ldots, b_n^*)$ is decidable as either holding or not holding in $N\|$;

(d) we define the witness $\mathcal{W}_{(N,\ Algorithmic)}$ as any simple functional language that gives evidence that $\|\text{SATCON}(\mathcal{I}_{PA(N)})\|$ is always *effectively* decidable in N:

Lemma 3 *$A^*(x_1, x_2, \ldots, x_n)$ is always algorithmically computable in N by $\mathcal{W}_{(N,\ Algorithmic)}$.*

Proof If $[A(x_1, x_2, \ldots, x_n)]$ is an atomic formula of PA then, for any given sequence of numerals $[b_1, b_2, \ldots, b_n]$, the PA formula $[A(b_1, b_2, \ldots, b_n)]$ is an atomic formula of the form $[c = d]$, where $[c]$ and $[d]$ are atomic PA formulas that denote PA numerals. Since $[c]$ and $[d]$ are recursively defined formulas in the language of PA, it follows from a standard result[40] that, if PA is consistent, then $[c = d]$ is algorithmically computable as either true or false in N. In other words, if PA is consistent, then $[A(x_1, x_2, \ldots, x_n)]$ is algorithmically computable (since there is an algorithm that, for any given sequence of numerals $[b_1, b_2, \ldots, b_n]$, will give evidence whether $[A(b_1, b_2, \ldots, b_n)]$ interprets as true or false in N. The lemma follows. □

It follows that $\mathcal{I}_{PA(N,\ Algorithmic)}$ is an algorithmic formulation of the 'standard' interpretation of PA over N in which we do not extraneously assume either that Aristotle's particularisation holds over N or, equivalently, that PA is ω-consistent.

5 Formally defining the standard interpretation of PA over N *constructively*

It follows from the analysis of the applicability of Tarski's inductive definitions of 'satisfiability' and 'truth' in Section 4 that we can formally define the standard interpretation $\mathcal{I}_{PA(N,\ Standard)}$ of PA *constructively* where:

(a) we define S as PA with standard first-order predicate calculus as the underlying logic;

(b) we define D as N;

(c) we take $\text{SM}(\mathcal{I}_{PA(N,\ Standard)})$ as any simple functional language.

We note that:

[40] For any natural numbers $m,\ n$, if $m \neq n$, then PA proves $[\neg(m = n)]$ ([Me64], p.110, Proposition 3.6). The converse is obviously true.

Theorem 2 *The atomic formulas of PA are algorithmically verifiable under the standard interpretation $\mathcal{I}_{PA(N,\ Standard)}$.*

Proof If $[A(x_1, x_2, \ldots, x_n)]$ is an atomic formula of PA then, for any given denumerable sequence of numerals $[b_1, b_2, \ldots]$, the PA formula $[A(b_1, b_2, \ldots, b_n)]$ is an atomic formula of the form $[c = d]$, where $[c]$ and $[d]$ are atomic PA formulas that denote PA numerals. Since $[c]$ and $[d]$ are recursively defined formulas in the language of PA, it follows from a standard result that, if PA is consistent, then $[c = d]$ interprets as the proposition $c = d$ which either holds or not for a witness $\mathcal{W}_N$ in N.

Hence, if PA is consistent, then $[A(x_1, x_2, \ldots, x_n)]$ is algorithmically verifiable since, for any given denumerable sequence of numerals $[b_1, b_2, \ldots]$, we can define an algorithm that provides evidence that the PA formula $[A(b_1, b_2, \ldots, b_n)]$ is decidable under the interpretation.

The theorem follows. □

It immediately follows that:

Corollary 1 *The 'satisfaction' and 'truth' of PA formulas containing logical constants can be defined under the standard interpretation of PA over N in terms of the evidence provided by the computations of a simple functional language.*

Corollary 2 *The PA-formulas are decidable under the standard interpretation of PA over N if, and only if, they are algorithmically verifiable under the interpretation.*

5.1 Defining 'algorithmic truth' under the standard interpretation of PA over N

Now we note that, in addition to Theorem 2:

Theorem 3 *The atomic formulas of PA are algorithmically computable under the standard interpretation $\mathcal{I}_{PA(N,\ Standard)}$.*

Proof If $[A(x_1, x_2, \ldots, x_n)]$ is an atomic formula of PA then we can define an algorithm that, for any given denumerable sequence of numerals $[b_1, b_2, \ldots]$, provides evidence whether the PA formula $[A(b_1, b_2, \ldots, b_n)]$ is true or false under the interpretation.

The theorem follows. □

This suggests the following definitions:

Definition 9 *A well-formed formula $[A]$ of PA is algorithmically true under $\mathcal{I}_{PA(N,\ Standard)}$ if, and only if, there is an algorithm which provides evidence that, given any denumerable sequence t of N, t satisfies $[A]$;*

Definition 10 *A well-formed formula $[A]$ of PA is algorithmically false under $\mathcal{I}_{PA(N,\ Standard)}$ if, and only if, it is not algorithmically true under $\mathcal{I}_{PA(N)}$.*

5.2 The PA axioms are algorithmically computable

The significance of defining 'algorithmic truth' under $\mathcal{I}_{PA(N,\ Standard)}$ as above is that:

Lemma 4 *The PA axioms PA_1 to PA_8 are algorithmically computable as algorithmically true over N under the interpretation $\mathcal{I}_{PA(N,\ Standard)}$.*

Proof Since $[x + y]$, $[x \star y]$, $[x = y]$, $[x']$ are defined recursively[41], the PA axioms PA_1 to PA_8 interpret as recursive relations that do not involve any quantification. The lemma follows straightforwardly from Definitions 3 to 8 in Section 4 and Theorem 2. □

Lemma 5 *For any given PA formula $[F(x)]$, the Induction axiom schema $[F(0) \rightarrow (((\forall x)(F(x) \rightarrow F(x')) \rightarrow (\forall x)F(x))]$ interprets as algorithmically true under $\mathcal{I}_{PA(N,\ Standard)}$.*

Proof By Definitions 3 to 10:

(a) If $[F(0)]$ interprets as algorithmically false under $\mathcal{I}_{PA(N,\ Standard)}$ the lemma is proved.

Since $[F(0) \rightarrow (((\forall x)(F(x) \rightarrow F(x')) \rightarrow (\forall x)F(x))]$ interprets as algorithmically true if, and only if, either $[F(0)]$ interprets as algorithmically false or $[((\forall x)(F(x) \rightarrow F(x')) \rightarrow (\forall x)F(x)]$ interprets as algorithmically true.

(b) If $[F(0)]$ interprets as algorithmically true and $[(\forall x)(F(x) \rightarrow F(x'))]$ interprets as algorithmically false under $\mathcal{I}_{PA(N,\ Standard)}$, the lemma is proved.

(c) If $[F(0)]$ and $[(\forall x)(F(x) \rightarrow F(x'))]$ both interpret as algorithmically true under $\mathcal{I}_{PA(N,\ Standard)}$, then by Definition 9 there is an algorithm which, for any natural number n, will give evidence that the formula $[F(n) \rightarrow F(n')]$ is true under $\mathcal{I}_{PA(N,\ Standard)}$.

Since $[F(0)]$ interprets as algorithmically true under $\mathcal{I}_{PA(N,\ Standard)}$, it follows that there is an algorithm which, for any natural number n, will give evidence that the formula $[F(n)]$ is true under the interpretation.

Hence $[(\forall x)F(x)]$ is algorithmically true under $\mathcal{I}_{PA(N,\ Standard)}$.

Since the above cases are exhaustive, the lemma follows. □

The Poincaré-Hilbert debate: We note that Lemma 5 appears to settle the Poincaré-Hilbert debate[42] in the latter's favour. Poincaré believed that the Induction Axiom could not be justified finitarily, as any such argument would necessarily need to appeal to infinite induction. Hilbert believed that a finitary proof of the consistency of PA was possible.

Lemma 6 *Generalisation preserves algorithmic truth under $\mathcal{I}_{PA(N,\ Standard)}$.*

Proof The two meta-assertions:

'$[F(x)]$ interprets as algorithmically true under $\mathcal{I}_{PA(N,\ Standard)}$'[43],

and

'$[(\forall x)F(x)]$ interprets as algorithmically true under $\mathcal{I}_{PA(N,\ Standard)}$'

both mean:

$[F(x)]$ is algorithmically computable as always true under $\mathcal{I}_{PA(N),\ Standard)}$. □

It is also straightforward to see that:

Lemma 7 *Modus Ponens preserves algorithmic truth under $\mathcal{I}_{PA(N,\ Standard)}$.* □

We thus have that:

Theorem 4 *The axioms of PA are always algorithmically true under the interpretation $\mathcal{I}_{PA(N,\ Standard)}$, and the rules of inference of PA preserve the properties of algorithmic satisfaction/truth under $\mathcal{I}_{PA(N,\ Standard)}$*[44]. □

5.3 The interpretation $\mathcal{I}_{PA(N,\ Algorithmic)}$ of PA over N is sound

We conclude from Section 4.3 and Section 5.2 that there is an algorithmic interpretation $\mathcal{I}_{PA(N,\ Algorithmic)}$ of PA over N such that:

Theorem 5 *The interpretation $\mathcal{I}_{PA(N,\ Algorithmic)}$ of PA is sound over N.*

Proof It follows immediately from Theorem 4 that the axioms of PA are always true under the interpretation $\mathcal{I}_{PA(N,\ Algorithmic)}$, and the rules of inference of PA preserve the properties of satisfaction/truth under $\mathcal{I}_{PA(N,\ Algorithmic)}$. □

We thus have a finitary proof that:

Theorem 6 *PA is consistent.* □

6 Conclusion

We have shown that although conventional wisdom is justified in *assuming* that the quantified arithmetical propositions of the first order Peano Arithmetic PA are *constructively* decidable under the standard interpretation of PA over the domain N of the natural numbers, the assumption does not address—and implicitly conceals—a significant ambiguity that needs to be made explicit.

Reason: Tarski's inductive definitions admit evidence-based interpretations of the first-order Peano Arithmetic PA that allow us to define the satisfaction and truth of the quantified formulas of PA *constructively* over N in *two* essentially different ways.

First in terms of algorithmic verifiabilty. We show that this allows us to define a *formal instantiational* interpretation $\mathcal{I}_{PA(\mathbb{N},\ Instantiational)}$ of PA over the domain $\mathbb{N}$ of the PA numerals that is sound (i.e. PA theorems interpret as true in N) if, and only if, the standard interpretation of PA over N—which is not known to be finitary—is sound.

Second in terms of algorithmic computability. We show that this allows us to define a finitary *algorithmic* interpretation $\mathcal{I}_{PA(N,\ Algorithmic)}$ of PA over N which *is* sound, and so we may conclude that PA is consistent.

---

[41] cf. [Go31], p.17.

[42] See [Hi27], p.472; also [Br13], p.59; [We27], p.482; [Pa71], p.502-503.

[43] See Definition 7

[44] Without appeal, moreover, to Aristotle's particularisation.

ACKNOWLEDGEMENTS

We would like to thank Professor Rohit Parikh for his suggestion that this paper should appeal to the computations of a simple functional language in general, and avoid appealing to the computations of a Turing machine in particular.

REFERENCES

[BBJ03] George S. Boolos, John P. Burgess, Richard C. Jeffrey. 2003. *Computability and Logic* (4th ed). Cambridge University Press, Cambridge.

[Be59] Evert W. Beth. 1959. *The Foundations of Mathematics*. Studies in Logic and the Foundations of Mathematics. Edited by L. E. J. Brouwer, E. W. Beth, A. Heyting. 1959. North Holland Publishing Company, Amsterdam.

[BF58] Paul Bernays and Abraham A. Fraenkel. 1958. *Axiomatic Set Theory*. Studies in Logic and the Foundations of Mathematics. Edited by L. E. J. Brouwer, E. W. Beth, A. Heyting. 1959. North Holland Publishing Company, Amsterdam.

[Br08] L. E. J. Brouwer. 1908. *The Unreliability of the Logical Principles*. English translation in A. Heyting, Ed. L. E. J. Brouwer: Collected Works 1: *Philosophy and Foundations of Mathematics*. Amsterdam: North Holland / New York: American Elsevier (1975): pp. 107-111.

[Br13] L. E. J. Brouwer. 1913. *Intuitionism and Formalism*. Inaugural address at the University of Amsterdam, October 14, 1912. Translated by Professor Arnold Dresden for the Bulletin of the American Mathematical Society, Volume 20 (1913), pp.81-96. 1999. Electronically published in Bulletin (New Series) of the American Mathematical Society, Volume 37, Number 1, pp.55-64.

[Co66] Paul J. Cohen. 1966. *Set Theory and the Continuum Hypothesis*. (Lecture notes given at Harvard University, Spring 1965) W. A. Benjamin, Inc., New York.

[Cr05] John N. Crossley. 2005. *What is Mathematical Logic? A Survey*. Address at the *First Indian Conference on Logic and its Relationship with Other Disciplines* held at the Indian Institute of Technology, Powai, Mumbai from January 8 to 12. Reprinted in *Logic at the Crossroads: An Interdisciplinary View - Volume I* (pp.3-18). ed. Amitabha Gupta, Rohit Parikh and Johan van Bentham. 2007. Allied Publishers Private Limited, Mumbai.

[Da82] Martin Davis. 1958. *Computability and Unsolvability*. 1982 ed. Dover Publications, Inc., New York.

[EC89] Richard L. Epstein, Walter A. Carnielli. 1989. *Computability: Computable Functions, Logic, and the Foundations of Mathematics*. Wadsworth & Brooks, California.

[Fe06] Solomon Feferman. 2006. *Are There Absolutely Unsolvable Problems? Gödel's Dichotomy*. Philosophia Mathematica (2006) 14 (2): 134-152.

[Fe08] Solomon Feferman. 2008. *Lieber Herr Bernays!, Lieber Herr Gödel! Gödel on finitism, constructivity and Hilbert's program*. in *Special Issue: Gödel's dialectica Interpretation* Dialectica, Volume 62, Issue 2, June 2008, pp. 245-290.

[Go31] Kurt Gödel. 1931. *On formally undecidable propositions of Principia Mathematica and related systems I*. Translated by Elliott Mendelson. In M. Davis (ed.). 1965. *The Undecidable*. Raven Press, New York.

[HA28] David Hilbert & Wilhelm Ackermann. 1928. *Principles of Mathematical Logic*. Translation of the second edition of the *Grundzüge Der Theoretischen Logik*. 1928. Springer, Berlin. 1950. Chelsea Publishing Company, New York.

[Hi25] David Hilbert. 1925. *On the Infinite*. Text of an address delivered in Münster on 4th June 1925 at a meeting of the Westphalian Mathematical Society. In Jean van Heijenoort. 1967. Ed. *From Frege to Gödel: A source book in Mathematical Logic, 1878 - 1931*. Harvard University Press, Cambridge, Massachusetts.

[Hi27] David Hilbert. 1927. *The Foundations of Mathematics*. Text of an address delivered in July 1927 at the Hamburg Mathematical Seminar. In Jean van Heijenoort. 1967. Ed. *From Frege to Gödel: A source book in Mathematical Logic, 1878 - 1931*. Harvard University Press, Cambridge, Massachusetts.

[Kl52] Stephen Cole Kleene. 1952. *Introduction to Metamathematics*. North Holland Publishing Company, Amsterdam.

[Kn63] G. T. Kneebone. 1963. *Mathematical Logic and the Foundations of Mathematics: An Introductory Survey*. D. Van Norstrand Company Limited, London.

[Li64] A. H. Lightstone. 1964. *The Axiomatic Method*. Prentice Hall, NJ.

[Me64] Elliott Mendelson. 1964. *Introduction to Mathematical Logic*. Van Norstrand, Princeton.

[Mu91] Chetan R. Murthy. 1991. *An Evaluation Semantics for Classical Proofs*. Proceedings of Sixth IEEE Symposium on Logic in Computer Science, pp. 96-109, (also Cornell TR 91-1213), 1991.

[Nv64] P. S. Novikov. 1964. *Elements of Mathematical Logic*. Oliver & Boyd, Edinburgh and London.

[Pa71] Rohit Parikh. 1971. *Existence and Feasibility in Arithmetic*. The Journal of Symbolic Logic, Vol.36, No. 3 (Sep., 1971), pp. 494-508.

[Qu63] Willard Van Orman Quine. 1963. *Set Theory and its Logic*. Harvard University Press, Cambridge, Massachusette.

[Rg87] Hartley Rogers Jr. 1987. *Theory of Recursive Functions and Effective Computability*. MIT Press, Cambridge, Massachusetts.

[Ro53] J. Barkley Rosser. 1953. *Logic for Mathematicians*. McGraw Hill, New York.

[Ru53] Walter Rudin. 1953. *Principles of Mathematical Analysis*. McGraw Hill, New York.

[Sh67] Joseph R. Shoenfield. 1967. *Mathematical Logic*. Reprinted 2001. A. K. Peters Ltd., Massachusetts.

[Sk28] Thoralf Skolem. 1928. *On Mathematical Logic*. Text of a lecture delivered on 22nd October 1928 before the Norwegian Mathematical Association. In Jean van Heijenoort. 1967. Ed. *From Frege to Gödel: A source book in Mathematical Logic, 1878 - 1931*. Harvard University Press, Cambridge, Massachusetts.

[Sm92] Raymond M. Smullyan. 1992. *Gödel's Incompleteness Theorems*. Oxford University Press, Inc., New York.

[Su60] Patrick Suppes. 1960. *Axiomatic Set Theory*. Van Norstrand, Princeton.

[Ta33] Alfred Tarski. 1933. *The concept of truth in the languages of the deductive sciences*. In *Logic, Semantics, Metamathematics, papers from 1923 to 1938* (p152-278). ed. John Corcoran. 1983. Hackett Publishing Company, Indianapolis.

[Tu36] Alan Turing. 1936. *On computable numbers, with an application to the Entscheidungsproblem*. In M. Davis (ed.). 1965. *The Undecidable*. Raven Press, New York. Reprinted from the Proceedings of the London Mathematical Society, ser. 2. vol. 42 (1936-7), pp.230-265; corrections, Ibid, vol 43 (1937) pp. 544-546.

[Wa63] Hao Wang. 1963. *A survey of Mathematical Logic*. North Holland Publishing Company, Amsterdam.

[We27] Hermann Weyl. 1927. *Comments on Hilbert's second lecture on the foundations of mathematics*. In Jean van Heijenoort. 1967. Ed. *From Frege to Gödel: A source book in Mathematical Logic, 1878 - 1931*. Harvard University Press, Cambridge, Massachusetts.

Machine Intention

Don Berkich [1]

Abstract. Skeptics find the thought that a robot can act of its own accord puzzling: Why should we think that a mere *artifact*, no matter how complicated, could ever have the capacity to act of its own accord given that its purpose and function is completely determined by its design specification? That is, why should we think that any such creation could be more than a mere cog in its causal environment? The skeptic's intuition is that machine agency is deeply incompatible with machine-hood in just the way it is not with person-hood. Thus the actions of, say, a situated robot like the Mars rovers cannot be more than a mere extension of the roboticist's agency inasmuch as the robot's design tethers it to the roboticist's intentions. In this talk I delve into the strongest version of the skeptical argument I've been able to make out so as to explore the roboticist's challenge.

Introduction

It can be argued that there exists a counterpart to the distinction between *original intentionality* and *derived intentionality* in agency: Given its design specification, a machine's agency is at most derived from its designer's original agency, even if the machine's resulting behavior sometimes surprises or dismays the designer. The argument for drawing this distinction hinges on the notion that intentions are necessarily *conferred* on machines by their designers' ambitions. To be sure, this is a decidedly negative conclusion for the prospects of strong artificial intelligence.

In this paper I wish to turn the tables by dismantling the strongest argument I can locate in the philosophical literature against the possibility of original machine agency, with the following caveat. The artificial intelligence (AI) crowd, if I may, tends to dismiss the philosophy of mind crowd's demands as unreasonable in light of the range of highly sophisticated behaviors currently demonstrated by the most advanced robotic systems. The mind crowd's objections, it is thought, result from an unfortunate lack of technical sophistication which leads to a failure to grasp the full import of the roboticists' achievements. The mind crowd's response is to point out that sophisticated behavior alone is not a sufficient condition on full-bore mentality: Thus John Searle's article in the February 23, 2011 issue of the Wall Street Journal is aptly entitled, "Watson Doesn't Know It Won on Jeopardy!" I think it a mistake for the AI crowd to dismiss the mind crowd's worries without very good reasons. By keeping the AI crowd's feet to the fire, the mind crowd is providing a welcome skeptical service. That said, sometimes there are very good reasons for the AI crowd to push back against the mind crowd; here I provide a specific and important case-in-point so as to illuminate some of the pitfalls in their complicated relationship.

In general, skeptical arguments against original machine agency may be stated in the *Modus Tollens* form:

1. If X is an original agent, then X must have property P.
2. No machine can have property P.
3. Therefore, no machine can be an original agent. 1&2

The force of each skeptical argument depends, of course, on the property P: The more clearly a given P is such as to be required by original agency but excluded by mechanism the better the skeptic's case. By locating property P in intention formation, Lynne Rudder Baker [2] identifies a particularly potent skeptical argument against original machine agency, if it succeeds. I proceed as follows. In the first section I set out and refine Baker's challenge. In the second section I propose a measured response. In the third and final section I use the measured response to draw attention to some of the excesses on both sides.

The Mind Crowd's Challenge: Baker's Skeptical Argument

Roughly put, Baker argues that machines cannot act since actions require intentions, intentions require a first-person perspective, and no amount of third-person information can bridge the gap to a first-person perspective. Baker [2, p. 157] usefully sets her own argument out:

Argument A

1. In order to be an agent, an entity must be able to formulate intentions.
2. In order to formulate intentions, an entity must have an irreducible first-person perspective.
3. Machines lack an irreducible first-person perspective.
4. Therfore, machines are not agents. 1,2&3

Baker has, however, failed to state her argument correctly. It is not just that machines are not (original) agents or do not happen presently to be agents, since that allows that at some point in the future machines may be agents or at least that machines can in principle be agents. Baker's conclusion is actually much stronger. As she outlines her own project, "[w]ithout denying that artificial models of intelligence may be useful for suggesting hypotheses to psychologists and neurophysiologists, I shall argue that there is a radical limitation to applying such models to human intelligence. And this limitation is exactly the reason why computers can't act." [2, p. 157]

Note that 'computers can't act' is substantially stronger than 'machines are not agents'. Baker wants to argue that it is impossible for machines to act, which is presumably more difficult than arguing that we don't at this time happen to have the technical sophistication to create machine agents. Revising Baker's extracted argument to bring it in line with her proposed conclusion, however, requires some corresponding strengthening of premise A.3, as follows:

[1] Texas AM University-Corpus Christi, Corpus Christi, Texas, email: don.berkich@tamucc.edu

Argument B

1. In order to be an original agent, an entity must be able to formulate intentions.
2. In order to formulate intentions, an entity must have an irreducible first-person perspective.
3. Machines necessarily lack an irreducible first-person perspective.
4. Therefore, machines cannot be original agents. 1,2&3

Argument B succeeds in capturing Baker's argument provided that her justification for B.3 has sufficient scope to conclude that machines cannot in principle have an irreducible first-person perspective. What support does she give for B.1, B.2, and B.3?

B.1 is true, Baker asserts, because original agency implies intentionality. She takes this to be virtually self-evident; the hallmark of original agency is the ability to form intentions, where intentions are to be understood on Castaneda's [4] model of being a "dispositional mental state of endorsingly thinking such thoughts as 'I shall do A'." [2, p. 157] B.2 and B.3, on the other hand, require an account of the first-person perspective such that

- The first person perspective is necessary for the ability to form intentions; and
- Machines necessarily lack it.

As Baker construes it, the first person perspective (FPP) has at least two essential properties. First, the FPP is irreducible, where the irreducibility in this case is due to a linguistic property of the words used to refer to persons. In particular, first person pronouns cannot be replaced with descriptions *salve veritate*. "First-person indicators are not simply substitutes for names or descriptions of ourselves." [2, p. 157] Thus Oedipus can, without absurdity, demand that the killer of Laius be found. "In short, thinking about oneself in the first-person way does not appear reducible to thinking about oneself in any other way." [2, p. 158]

Second, the FPP is necessary for the ability to "conceive of one's thoughts as one's own." [2, p. 158] Baker calls this 'second-order consciousness'. Thus, "if X cannot make first-person reference, then X may be conscious of the contents of his own thoughts, but not conscious that they are his own." [2, p. 158] In such a case, X fails to have second-order consciousness. It follows that "an entity which can think of propositions at all enjoys self-consciousness if and only if he can make irreducible first-person reference." [2, p. 158] Since the ability to form intentions is understood on Castaneda's model as the ability to endorsingly think propositions such as "I shall do A", and since such propositions essentially involve first-person reference, it is clear why the first person perspective is necessary for the ability to form intentions. So we have some reason to think that B.2 is true. But, apropos B.3, why should we think that machines necessarily lack the first-person perspective?

Baker's justification for B.3 is captured by her claim that "[c]omputers cannot make the same kind of reference to themselves that self-conscious beings make, and this difference points to a fundamental difference between humans and computers–namely, that humans, but not computers, have an irreducible first-person perspective." [2, p. 159] To make the case that computers are necessarily handicapped in that they cannot refer to themselves in the same way that self-conscious entities do, she invites us to consider what would have to be the case for a first person perspective to be programmable:

a) FPP can be the result of information processing.
b) First-person episodes can be the result of transformations on discrete input via specifiable rules. [2, p. 159]

Machines necessarily lack an irreducible first-person perspective since both (a) and (b) are false. (b) is straightforwardly false, since "the world we dwell in cannot be represented as some number of independent facts ordered by formalizable rules." [2, p. 160] Worse, (a) is false since it presupposes that the FPP can be generated by a rule governed process, yet the FPP "is not the result of any rule-governed process." [2, p. 160] That is to say, "no amount of third-person information about oneself ever compels a shift to first person knowledge." [2, p. 160] Although Baker does not explain what she means by "third-person information" and "first person knowledge," the point, presumably, is that there is an unbridgeable gap between the third-person statements and the first-person statements presupposed by the FPP. Yet since the possibility of an FPP being the result of information processing depends on bridging this gap, it follows that the FPP cannot be the result of information processing. Hence it is impossible for machines, having only the resource of information processing as they do, to have an irreducible first-person perspective.

Baker's skeptical challenge to the AI crowd may be set out in detail as follows:

Argument C

1. Necessarily, X is an original agent only if X has the capacity to formulate intentions.
2. Necessarily, X has the capacity to formulate intentions only if X has an irreducible first person perspective.
3. Necessarily, X has an irreducible first person perspective only if X has second-order consciousness.
4. Necessarily, X has second-order consciousness only if X has self-consciousness.
5. Therefore, necessarily, X is an original agent only if X has self-consciousness 1,2,3&4
6. Necessarily, X is a machine only if X is designed and programmed.
7. Necessarily, X is designed and programmed only if X operates just according to rule-governed transformations on discrete input.
8. Necessarily, X operates just according to rule-governed transformations on discrete input only if X lacks self-consciousness.
9. Therefore, necessarily, X is a machine only if X lacks self-consciousness. 6,7&8
10. Therefore, necessarily, X is a machine only if X is not an original agent. 5&9

A Measured Response on Behalf of the AI Crowd

While there presumably exist skeptical challenges which ought not be taken seriously because they are, for want of careful argumentation, themselves unserious, I submit that Baker's skeptical challenge to the AI crowd is serious and ought to be taken as such. It calls for a measured response. It would be a mistake, in other words, for the AI crowd to dismiss Baker's challenge out of hand for want of technical sophistication, say, in the absence of decisive counterarguments. Moreover, counterarguments will not be decisive if they simply ignore the underlying import of the skeptic's claims.

For example, given the weight of argument against physicalist solutions to the hard problem of consciousness generally, it would be incautious of the AI crowd to respond by rejecting C.8 (but see [5] for a comprehensive review of the hard problem). In simple terms, the AI crowd should join the mind crowd in finding it daft at this point for a roboticist to claim that *there is something it is like to be her robot*, however impressive the robot or resourceful the roboticist in building it.

A more modest strategy is to sidestep the hard problem of consciousness altogether by arguing that having an irreducible FPP is not, contrary to C.2, a necessary condition on the capacity to form intentions. This is the appropriate point to press provided that it also appeals to the mind crowd's own concerns. For instance, if it can be argued that the requirement of an irreducible FPP is too onerous even for persons to formulate intentions under ordinary circumstances, then Baker's assumption of Castaneda's account will be vulnerable to criticism from both sides. Working from the other direction, it must also be argued the notion of programming that justifies C.7 and C.8 is far too narrow even if we grant that programming an irreducible FPP is beyond our present abilities. The measured response I am presenting thus seeks to moderate the mind crowd's excessively demanding conception of intention while expanding their conception of programming so as to reconcile, in principle, the *prima facie* absurdity of a programmed (machine) intention.

Baker's proposal that the ability to form intentions implies an irreducible FPP is driven by her adoption of Castaneda's [4] analysis of intention: To formulate an intention to A is to endorsingly think the thought, "I shall do A". There are, however, other analyses of intention which avoid the requirement of an irreducible FPP. Davidson [6] sketches an analysis of what it is to form an intention to act: "an action is performed with a certain intention if it is caused in the right way by attitudes and beliefs that rationalize it." [6, p. 87] Thus,

> If someone performs an action of type A with the intention of performing an action of type B, then he must have a pro-attitude toward actions of type B (which may be expressed in the form: an action of type B is good (or has some other positive attribute)) and a belief that in performing an action of type A he will be (or probably will be) performing an action of type B (the belief may be expressed in the obvious way). The expressions of the belief and desire entail that actions of type A are, or probably will be, good (or desirable, just, dutiful, etc.). [6, pp. 86-87]

Davidson is proposing that S A's with the intention of B-ing only if

i. S has pro-attitudes towards actions of type B.
ii. S believes that by A-ing S will thereby B.

The pro-attitudes and beliefs S has which rationalize his action cause his action. But, of course, it is not the case that S's having pro-attitudes towards actions of type B and S's believing that by A-ing she will thereby B jointly implies that S actually A's with the intention of B-ing. (i) and (ii), in simpler terms, do not jointly suffice for S's A-ing with the intention of B-ing since it must be that S A's because of her pro-attitudes and beliefs. For Davidson, 'because' should be read in its causal sense. Reasons consisting as they do of pro-attitudes and beliefs cause the actions they rationalize.

Causation alone is not enough, however. To suffice for intentional action reasons must cause the action in the right way. Suppose (cf [6, pp. 84-85]) Smith gets on the plane marked 'London' with the intention of flying to London, England. Without alarm and without Smith's knowledge, a shy hijacker diverts the plane from its London, Ontario destination to London, England. Smith's beliefs and pro-attitudes caused him to get on the plane marked 'London' so as to fly to London, England. Smith's intention is satisfied, but only by accident, as it were. So it must be that Smith's reasons cause his action in the right way, thereby avoiding so called wayward causal chains. Hence, S A's with the intention of B-ing if, and only if,

i. S has pro-attitudes towards actions of type B.
ii. S believes that by A-ing S will thereby B.
iii. S's relevant pro-attitudes and beliefs cause her A-ing with the intention of B-ing in the right way.

Notice that there is no reference whatsoever involving an irreducible FPP in Davidson's account. Unlike Castaneda's account, there is no explicit mention of the first person indexical. So were it the case that Davidson thought animals could have beliefs, which he does not [7], it would be appropriate to conclude from Davidson's account that animals can act intentionally despite worries that animals would lack an irreducible first-person perspective. Presumably robots would not be far behind.

It is nevertheless open to Baker to ask about (ii): S believes that by A-ing S will thereby B. Even if S does not have to explicitly and endorsingly think, "I shall do A" to A intentionally, (ii) requires that S has a self-referential belief that by A-ing he himself will thereby B. Baker can gain purchase on the problem by pointing out that such a belief presupposes self-consciousness every bit as irreducible as the FPP.

Consider, however, that a necessary condition on Davidson's account of intentional action is that S believes that by A-ing S will thereby B. Must we, however, take 'S' in S's belief that by A-ing S will thereby B *de dicto*? Just as well, could it not be the case (*de re*) that S believes, of itself, that by A-ing it will thereby B?

The difference is important. Taken *de dicto*, S's belief presupposes self-consciousness since S's belief is equivalent to having the belief, "by A-ing I will thereby B". Taken (*de re*), however, S's belief presupposes at most self-*representation*, which can be tokened without solving the problem of (self) consciousness.

Indeed, it does not seem to be the case that the intentions I form presuppose either endorsingly thinking "I shall do A!" as Castaneda (and Baker) would have it or a *de dicto* belief that by A-ing I will B as Davidson would have it. Intention-formation is transparent: I simply believe that A-ing B's, so I A. The insertion of self-consciousness as an intermediary requirement in intention formation would effectively eliminate many intentions in light of environmental pressures to act quickly. Were Thog the caveman required to endorsingly think "I shall climb this tree to avoid the saber-toothed tiger" before scrambling up the tree he would lose precious seconds and, very likely, his life. Complexity, particularly temporal complexity, constrains us as much as it does any putative original machine agent. A theory of intention which avoids this trouble surely has the advantage over theories of intention which do not.

The mind crowd may nevertheless argue that even a suitably attenuated conception of intention cannot be programmed under Baker's conception of programming. What is her conception of programming? Recall that Baker defends B.3 by arguing that machines cannot achieve a first-person perspective since machines gain information *only* through rule-based transformations on discrete input and no amount or combination of such transformations could suffice for the transition from a third-person perspective to a first-person perspective. That is,

Argument D

1. If machines were able to have a FPP, then the FPP can be the result of transformations on discrete input via specifiable rules.
2. If the FPP can be the result of transformations on discrete input via specifiable rules, then there exists some amount of third-person information which compels a shift to first-person knowledge.

3. No amount of third-person information compels a shift to first-person knowledge.
4. Therefore, first-person episodes cannot be the result of transformations on discrete input via specifiable rules. 2&3
5. Therefore, machines necessarily lack an irreducible first-person perspective. 1&4

The problem with D is that it betrays an overly narrow conception of machines and programming, and this is true even if we grant that we don't presently know of *any* programming strategy that would bring about an irreducible FPP.

Here is a simple way of thinking about machines and programming as D would have it. There was at one time (for all I know, there may still be) a child's toy which was essentially a wind-up car. The car came with a series of small plastic disks, with notches around the circumference, which could be fitted over a rotating spindle in the middle of the car. The disks acted as a cam, actuating a lever which turned the wheels when the lever hit a notch in the side of the disk. Each disk had a distinct pattern of notches and resulted in a distinct route. Thus, placing a particular disk on the car's spindle 'programs' the car to follow a particular route.

Insofar as it requires that programming be restricted to transformations on discrete input via specifiable rules, Argument D treats all machines as strictly analogous to the toy car and programming as analogous to carving out new notches on a disk used in the toy car. Certainly Argument D allows for machines which are much more complicated than the toy car, but the basic relationship between program and machine behavior is the same throughout. The program determines the machine's behavior, while the program itself is in turn determined by the programmer. It is the point of D.2 that, if an irreducible FPP were programmable, it would have to be because the third-person information which can be supplied by the programmer suffices for a first-person perspective, since all the machine has access to is what can be supplied by a programmer.

Why should we think that a machine's only source of information is what the programmer provides? Here are a few reasons to think that machines are not so restricted:

- Given appropriate sensory modalities and appropriate recognition routines, machines are able to gain information about their environment without that information having been programmed in advance. [1] It would be as if the toy car had an echo-locator on the front and a controlling disk which notched itself in reaction to obstacles so as to maneuver around them.
- Machines can be so constructed as to 'learn' by a variety of techniques. [8] Even classical conditioning techniques have been used. The point is merely that suitably constructed, a machine can put together information about its environment and itself which is not coded in advance by the programmer and which is not available other than by, for example, trial and error. It would be as if the toy car had a navigation goal and could adjust the notches in its disk according to whether it is closer or farther from its goal.
- Machines can evolve. [3] Programs evolve through a process of mutation and extinction. Code in the form of so-called genetic algorithms is replicated and mutated. Unsuccessful mutations are culled, while successful algorithms are used as the basis for the next generation. Using this method one can develop a program for performing a particular task without having any knowledge of how the program goes about performing the task. Strictly speaking, there is no programmer for such programs. Here the analogy with the toy car breaks down somewhat. It's as if the toy car started out with a series of disks of differing notch configurations and the car can take a disk and either throw it out or use it as a template for further disks, depending on whether or not a given disk results in the car being stuck against an obstacle, for instance.
- Programs can be written which write their own programs. [3] A program can spawn an indefinite number of programs, including an exact copy of itself. It need not be the case that the programmer be able to predict what future code will be generated, since that code may be partially the result of information the machine gathers, via sensory modalities, from its environment. So, again, in a real sense there is no programmer for these programs. The toy car in this case starts out with a disk which itself generates disks and these disks may incorporate information about obstacles and pathways.

Indeed, many of the above techniques develop Turing's own suggestions:

> Let us return for a moment to Lady Lovelace's objection, which stated that the machine can only do what we tell it to do...
> Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child brain is something like a notebook as one buys it from the stationer's. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.) Our hope is that there is so little mechanism in the child brain that something like it can be easily programmed. The amount of work in the education we can assume, as a first approximation, to be much the same as for the human child.
> We have thus divided our problem into two parts. The child programme and the education process. These two remain very closely connected. We cannot expect to find a good child machine at the first attempt. One must experiment with teaching one such machine and see how well it learns...
> The idea of a learning machine may appear paradoxical to some readers. How can the rules of operation of the machine change? They should describe completely how the machine will react whatever its history might be, whatever changes it might undergo. The rules are thus quite time-invariant. This is quite true. The explanation of the paradox is that the rules which get changed in the learning process are of a rather less pretentious kind, claiming only an ephemeral validity. The reader may draw a parallel with the Constitution of the United States. [9, pp. 454-458]

As Turing anticipated, machines can have access to information and utilize it in ways which are completely beyond the purview of the programmer. So while it may not be the case that a programmer can write code for an irreducible FPP, as Argument D requires, it still can be argued that the sources of information available to a suitably programmed robot nevertheless enable it to formulate intentions when intentions do not also presuppose an irreducible FPP.

Consider the spectacularly successful Mars rovers Spirit and Opportunity. Although the larger goal of moving from one location to another was provided by mission control, specific routes were determined *in situ* by constructing maps and evaluating plausible routes according to obstacles, inclines, etc. Thus the Mars rovers were, in a rudimentary sense, gleaning information from their environment and using that information to assess alternatives so as to plan and execute subsequent actions. None of this was done with the requirement of, or pretense to having, an irreducible FPP, yet it does come

closer to fitting the Davidsonian model of intentions. To be sure, this is intention-formation of the crudest sort, and it requires further argument that propositional attitudes themselves are computationally tractable.

A Larger Point: Avoiding Excesses on Both Sides

Baker closes by pointing out that robots' putative inability to form intentions has far-reaching implications:

> So machines cannot engage in intentional behavior of any kind. For example, they cannot tell lies, since lying involves the intent to deceive; they cannot try to avoid mistakes, since trying to avoid mistakes entails intending to conform to some normative rule. They cannot be malevolent, since having no intentions at all, they can hardly have wicked intentions. And, most significantly, computers cannot use language to make assertions, ask questions, or make promises, etc., since speech acts are but a species of intentional action. Thus, we may conclude that a computer can never have a will of its own. [2, p. 163]

The challenge for the AI crowd, then, is to break the link Baker insists exists between intention formation and an irreducible FPP. For if Baker is correct and the FPP presupposes self-consciousness, the only way the roboticist can secure machine agency is by solving the vastly more difficult problem of consciousness, which so far as we presently know is computationally intractable. I have argued that the link can be broken, provided a defensible and computationally tractable account of intention is available to replace Castaneda's overly demanding account.

If my analysis is sound, then there are times when it is appropriate for the AI crowd to push back against the mind crowd. Yet they must do so in such a way as to respect so far as possible the ordinary notions the mind crowd expects to see employed. In this case, were the AI crowd to so distort the concept of intention in their use of the term that it no longer meets the mind crowd's best expectations, the AI crowd would merely have supplied the mind crowd with further skeptical arguments. In this sense, the mind crowd plays a valuable role in demanding that the AI crowd ground their efforts in justifiable conceptual requirements, which in no way entails that the AI crowd need accept those conceptual requirements without further argument. Thus the enterprise of artificial intelligence has as much to do with illuminating the efforts of the philosophers of mind as the latter have in informing those working in artificial intelligence.

This is a plea by example, then, to the AI crowd that they avoid being overly satisfied with themselves simply for simulating interesting behaviors, unless of course the point of the simulation is the behavior. At the same time, it is a plea to the mind crowd that they recognize when their claims go too far even for human agents and realize that the AI crowd is constantly adding to their repertoire techniques which can and should inform efforts in the philosophy of mind.

ACKNOWLEDGEMENTS

I am grateful to two anonymous referees for their helpful comments and criticisms.

REFERENCES

[1] R.C. Arkin, *Behavior Based Robotics*, MIT Press, Cambridge, Mass., 1998.

[2] L.R. Baker, 'Pwhy computer's can't act', *American Philosophical Quarterly*, **18**, 157–163, (1981).

[3] D. H. Ballard, *An Introduction to Natural Computation*, MIT Press, Cambridge, Mass., 1997.

[4] H-N. Castaneda, *Thinking and Doing: The Philosophical Foundations of Institutions*, D. Reidel Publishing Co., Dordrecht, 1975.

[5] D. Chalmers, *Consciousness and Its Place in Nature*, 247–272, Philosophy of Mind: Classical and Contemporary Readings, Oxford University Press, Oxford, 2002.

[6] D. Davidson, *Intending*, 83–102, Essays on Actions and Events, Clarendon Press, Oxford, 1980.

[7] D. Davidson, *Thought and Talk*, 155–170, Inquiries into Truth and Interpretation, Clarendon Press, Oxford, 1984.

[8] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press. A Bradford Book, Cambridge, Mass., 3rd edn., 1998.

[9] A.M. Turing, 'Computing machinery and intelligence', *Mind*, **59**, 433–60, (1950).

A lesson from subjective computing: autonomous self-referentiality and social interaction as conditions for subjectivity

Patrick Grüneberg[1] and **Kenji Suzuki**[2]

Abstract. In this paper, we model a relational notion of subjectivity by means of two experiments in subjective computing. The goal is to determine to what extent a cognitive and social robot can be regarded to act subjectively. The system was implemented as a reinforcement learning agent with a coaching function. To analyze the robotic agent we used the method of levels of abstraction in order to analyze the agent at four levels of abstraction. At one level the agent is described in mentalistic or subjective language respectively. By mapping this mentalistic to an algorithmic, functional, and relational level, we can show to what extent the agent behaves subjectively as we make use of a relational concept of subjectivity that draws upon the relations that hold between the agent and its environment. According to a relational notion of subjectivity, an agent is supposed to be subjective if it exhibits autonomous relations to itself and others, i.e. the agent is not fully determined by a given input but is able to operate on its input and decide what to do with it. This theoretical notion is confirmed by the technical implementation of self-referentiality and social interaction in that the agent shows improved behavior compared to agents without the ability of subjective computing. On the one hand, a relational concept of subjectivity is confirmed, whereas on the other hand, the technical framework of subjective computing is being theoretically founded.

1 INTRODUCTION

The mental phenomenon called 'subjectivity' has been up to present days one of the central topics of philosophical discussion. Even before the proclamation of the 'subject' as a principal of knowledge, one might regard the relation of an epistemic and practical agent to the world and itself as one of the most notorious issues even in antique and medieval philosophy. However, in these days, 'subjectivity' enjoys great popularity as phenomenal consciousness, as the individual first-person perspective.[3] But 'subjectivity' needs not necessarily be related to consciousness. Instead, recent developments in robotics show that 'subjectivity' can also be related to intelligence. Actually, the idea to analyze 'subjectivity' as intelligence is not that new [8], [9]. One obvious advantage of decoupling 'subjectivity' from consciousness is that intelligence can be analyzed without making use of the most difficult concepts of (phenomenal) consciousness. From this perspective, 'subjectivity' is conceptualized as a relational concept, i.e. subjectivity comprises certain relations of an agent towards itself and its environment [16]. The question of phenomenal consciousness is then subordinated in favor of the agent's self-relation and relations to others. An agent then is supposed to be subjective if it exhibits autonomous relations to itself and others, i.e. the agent is not fully determined by a given input but is able to operate on its input and decide what to do with it. This relational perspective also allows us to take into account social relations. Accordingly, intelligence is not solely a product of internal processes but is constituted in the course of social interaction and therefore builds on the notions of *Aufforderung* [2] and recognition [18].

To narrow down and as an attempt to verify this philosophical and quite abstract notion of subjectivity, we refer to two experiments in subjective computing [17], [14]. Subjective computing aims at utilizing insights from human subjectivity in general, the perceptual process, and human-human interaction for the design of algorithms and human-robot interaction (this concept was initially proposed in [21]). From an engineering perspective it is not the goal to give an account on subjectivity, but rather to utilize certain aspects of human cognition to solve specific problems. One major problem concerns reinforcement learning (RL). Even if an agent is able to decide autonomously about the modification of its behavior, the agent still has to learn what kind of behavior is well suited in order to accomplish a certain task. In order to evaluate the agent's behavior a coaching function has been implemented into a RL agent so that the agent can receive a trainer's feedback. The crucial point regarding the potential subjectivity of this agent is that this feedback does not modify the agent's behavior directly, but the agent interprets the feedback and decides about the subsequent modification of its behavior by itself. Thus, with regard to the relational notion of subjectivity, the agent relates to itself while interpreting feedback and at the same time socially relates to a human trainer. For the implementation of this robotic system both relations, the self-relation in the course of interpretation and the relation to others (the human trainer) enable the robotic agent to successfully accomplish a difficult learning task.

In our relational analysis of the robotic agent we draw upon the ascription of mental abilities to the robotic agent that is supposed to observe, interpret, and reflect the feedback. The question is to what extent this mental behavior is finally algorithmically implemented. By means of an analysis of levels of abstraction [10], we relate the algorithmic to the mentalistic level. The focus lies on an intermediate relational level where it can be shown that the robotic agent exhibits an autonomous self-relation and a social relation to others.

[1] Artficial Intelligence Laboratory, University of Tsukuba, Japan; Institut für Philosophie, Literatur-, Wissenschafts- und Technikgeschichte, Technical University Berlin, Germany, e-mail: patrick.grueneberg@tu-berlin.de.

[2] Center for Cybernics Research, University of Tsukuba, Japan, Japan Science and Technology Agency, Japan, e-mail: kenji@ieee.org.

[3] [24] and [19] are just two of the most prominent examples for phenomenally based accounts. [22] follows a cognitive science approach to subjectivity in terms of a phenomenal perspective to the world and the agent itself.

Even if the robotic agent cannot be conceived of as a full-blown subject (compared to humans), the successful implementation of autonomous self-referentiality and a social relation to others allows us to ascribe subjective states to the robotic agent. Even if the robotic agent cannot be regarded as a full-blown human subject, the successful engineering implementation of this relational structure can be seen as a confirmation for the philosophical notion of subjectivity as increased intelligent behavior has been gained.

In section 2 we will start by shortly introducing the relational concept of subjectivity and by explaining our case study, "coaching a robot based on human affective feedback". After describing the algorithm and the coaching function as well as the experimental layout and results, in section 3, we introduce Floridi's method of levels of abstraction by explaining what this method consists in and how we use it to analyze the robotic agent. By means of four levels of abstractions we will then analyze the robotic agent with a focus on its relational structure. Section 4 begins with an informal analysis that is followed by a formal treatment of the levels of abstraction and their relations. Finally, we evaluate to what extent mental subjective abilities can be ascribed to the robotic agent.

2 PHILOSOPHICAL AND TECHNICAL BACKGROUNDS

2.1 Relational subjectivity

To introduce the concept of relational subjectivity, it is helpful to refer to the current philosophical debate, especially to phenomenal and first-person accounts of subjectivity which enjoy great popularity (see note 3). According to the general idea of this framework, subjectivity consists in phenomenal consciousness so that a phenomenal subject is able to experience its mental states. These mental states refer to environmental entities (or their features, respectively) or states of the agent itself. The experience of these states is subjective to the extent that this experience is only accessible for the agent who has it. Furthermore, such accounts are often representationally or, at least, realistically based, i.e. the experience refers to an objectively existing (independently of the agent) world that becomes conscious in the agent's mind. Although there are plenty of different versions of phenomenal and first-person, representational and realistic accounts, the crucial point for our investigation lies in decoupling subjectivity from the any kind of phenomenal consciousness (for further explanation of the methodological arguments for this decoupling see [16]).

Instead, subjectivity can be grounded in action. In a Kantian and therefore transcendental perspective, this action is conceived of as a condition of the possibility of subjectivity. The main purpose of this action is to structure and construct the subject's reality by means of schematic capacities.[4]. These schematic capacities generate the subject's attitude towards a given reality in which the subject can act. Hence, subjectivity has, secondly, to be decoupled from the notion of a *psychological* subject. The distinction between different individual subjects is not based on different individuals. Instead, the schematic processes make the difference as these exhibit necessary features that apply for every individual subject. Accordingly, subjects usually share the same space-time dimension. On the other hand, schematic processes are not completely determined and thus allow for voluntary action that depends on individual decisions, e.g. on the individual use of cognitive abilities as perception or action. An individual subject can voluntarily focus its visual attention to a certain position in space and decide to move in this direction or to hold its current place. In turn, these voluntary actions depend on determinations that are out of reach for the individual subject, i.e. when visual attention has been focused to a certain position, then the content of the visual experience is determined.

Accordingly, subjectivity is relationally generated by simultaneous processes of determining and voluntary schematic activity. One and the same cognitive action underlies this twofold schematism so that subjectivity is conceived as a relational momentum that is generated in opposition to an objective or determining momentum in an agent's information space. This twofold structure also applies for the individual agent that acts in the social context of other individual agents: On the one hand, the agent relies on its autonomous capacities. At the same time, it depends on social interaction as social interaction constrains its autonomy and therefore provokes a reaction. A reaction here is understood as a self-determination of the agent's actions provoked by some external constraint. Again, a subjective agent is conceived of as relationally constituted. This mutual interdependency of voluntarily determining and necessarily being determined forms the basic framework for a relational concept of subjectivity. In the following we are going to investigate two experiments in cognitive and social robotics in order to evaluate if and to what extent this relational concept of subjectivity can be computationally modeled and implemented. This serves to narrow down and concretize the quite abstract relational notion; at the same time, the framework of subjective computing can be made more explicit; especially we hope to clarify what it can mean for a robotic agent to behave subjectively.

2.2 Case study: coaching a robot based on human affective feedback

Generally, interaction and henceforth social intelligence are regarded as a constitutive part of intelligence at all [5]. Based on an interactive learning algorithm reciprocal interaction between a robotic agent and a human instructor is facilitated. This way of situated learning enables the coach to scaffolding acts of providing feedback [23], while the robot demonstrates its mastery of the task continuously by means of improved behavior. In this kind of peer-to-peer human-robot interaction the robotic agent has to perceive emotions and learn models of its human counterpart [11]. Hence the robot needs to be at least socially receptive, i.e. socially passive in order to benefit from interaction, or coaching respectively [1], and socially embedded, i.e. situated in a social environment and interacting with humans. If the agent is structurally coupled with the social environment, he will be able to be partially aware of human interactional structures [7]. In order to socialize robots have to be compatible with human's ways of interacting and communicating. On the other hand humans must be able to rely on the robot's actions and be allowed to have realistic expectations about its behavior.

In the context of embodied cognition, we are able to model subjectivity as an interactional (social) and therefore relational issue. This means that subjectivity is realized in the course of social interaction which is investigated in the field of social robotics. One core issue in designing social robots consists in socially situated learning. New skills or knowledge are acquired by interacting with other agents. Beside robot-robot interaction (so-called "swarm intelligence" or "collective intelligence"), human-robot interaction displays another major approach [6], [12]. We focus on the case of teaching a robot [28], [29] by means of coaching. Unlike teaching the coaching process does not depend on an "omniscient" teacher that guides the agent toward the goal, but the instructor only gives hints and clues in terms

[4] See historically [20] and, in the sense of an extended schematism, [9]; in the following we refer to an updated schematic account in [16]

of a binary feedback, i.e. positive or negative. It is then the robot's cognitive task to process this feedback and control its actions autonomously.

Our approach to subjective computing is based on two experiments on coaching a robot. These experiments were conducted at the Artificial Intelligence Laboratory (University of Tsukuba) previously to this investigation. The coaching process itself bears on two relational aspects that are the focus in these experiments:

1. the cognitive process of autonomous interpretation of the feedback by the agent [17]
2. the social interaction between the human instructor and the robot [14]

In the following we will, firstly, describe the problem that underlies the implementation of the coaching RL agent and of affective feedback, respectively. Secondly, we illustrate the experimental setups and results.

The first experiment [17] was conducted by Hirokawa and Suzuki and consists in a reinforcement learning (RL) agent with an implemented coaching function so that the robotic agent is open to human feedback during its behavior learning. While coaching had already been implemented before [25], [26], RL offers a significant advantage. A coaching RL agent is able to learn automatcially by its own internal values. RL is a commonly used method for autonomous machine learning based on the idea that an agent autonomously adapts to the specific constraints of an environment [27]. While often a learning algorithm is predefined regarding the parameters of an environment, an RL agent is able to adjust its learning process continuously during acting. This is done by continuously updating the expected reward of an action (state-value) by means of a reward function. The agent learns automatically when it conducts an action that matches the reward function and can subsequently shape its behavior in order to increase future rewards. The feature that is most relevant for our analyses is that the reward function defines which action can count as a successful action and therefore as a learning progress.

Yet, one central problem consists in the initial reward as the RL agent has to exploit a state space randomly by trial and error in order to discover the first reward. To avoid a time-consuming random search the reward function has to be carefully designed. However, this limits the flexibility of the algorithm. In order to bypass an exclusively trial-and-error search or a complicated design process, coaching is implemented in the RL agent by adding an interface to the RL agent that allocates a feedback signal [17]. RL then allows for coaching in that the human trainer gives feedback, and the learning agent adjusts its reward function and its action rules according to the feedback. Thus, the behavior is not directly instructed or trained, but the robot modifies its behavior by itself. At the same time the reward function does not need to be designed in advance. This autonomous estimation of the reward function then complements the standard RL based on a direct interaction with the environment.

In the experiment an RL agent controls a robotic arm in order to swing up and keep an inverted pendulum balanced. While carrying out the task, the RL agent receives continuously feedback in terms of human subjective cues, i.e. positive or negative [29]. The agent has to interpret this feedback and adjusts the reward function and therefore its actions accordingly. Thus, learning the reward function is based on simple and abstract (binary) feedback that is delivered in social interaction. The feedback itself does not determine the reward function directly, but allows the robot to modify the latter based on an act of interpretation that consists in an estimation of the input's relevancy to its own behavior. This interpretation depends on two successive criteria. Firstly, in contingency or causality detection the "agent determines specific states [of its behavior] that motivated the trainer to give feedback" ([17], p. 5), i.e. the agent identifies the feedback's target behavior that depends on a certain time range and a subsequent time delay specifying the time between the action and the feedback. This identification of target behavior is, secondly, complemented by a consistency or error detection, i.e. checking to what extent a given evaluation corresponds "to current and previous feedback to a similar behavior" ([17], p. 5f.). If the feedback is inconsistent (contradictory), it is regarded as irrelevant and the reward function will not be updated. In short, after assigning the feedback to a previous action and verifying its consistency the evaluation function is updated and action rules modified accordingly. In this way the robot exhibits an internal and manipulable model of the trainer's evaluation instead of just executing correction commands. Hence, different kinds of feedback (coaching strategies) lead to different degrees of rates of learning and success.

The second experiment [14] was conducted by Gruebler, Berenz, and Suzuki. At first it has to be noted that we draw on the second experiment in order to exemplify the significance of social interaction while the behaviorand learning algorithm differs from the RL agent in the first experiment. However, due to the binary feedback in both experiments the results can be complemented in the subsequent investigation of subjective relations. Hence, the second experiment concerns the it allocation of feedback [14]. Human feedback is delivered as a cue based on a binary (positive or negative) signal that is interpreted as confirmation or correction. Continuous non-verbal social cues are used as instructive input to help a humanoid robot to modify its behavior. While the robot is conducting a task, the human coach gives continuous feedback by means of smiling or frowning. The facial expression was measured by a wearable device that recognizes these basal facial movements as expressions of confirmation (smile) and correction (frown) [15]. In this way a binary feedback resulted that enabled the robot to modify its behavior continuously whilst conducting a task. No further specification of the signal is necessary. In this way the robotic agent is open to human affective feedback in direct interaction. The cognitive and interactional implementations of both experiments can be complemented to that effect that a binary signal is sufficient to instruct a robot while at the same time this signal can be allocated in a way very natural for humans.

2.3 Experimental layouts and results

Experiments on coaching a robot based on human subjective feedback form the ground for an analysis of a subjective agent. Both experimental setups that were introduced in the previously, are cases of HRI. The RL agent of the first experiment [17] has been implemented in a simulated and a real robotic arm whose learning task consisted in swinging up and keeping an inverted pendulum balanced (see Fig. 1). Instead of predesigning the reward function, the human instructor assists the RL agent by observing its behavior and giving a binary (positive or negative) feedback. In the real and the simulational setup a "significant improvement compared to the conventional RL with the same reward function" ([17], p. 14) had been measured as the conventional RL completely failed to achieve the task. The simulational setup additionally showed that the RL agent reflects coaching strategies of different instructors in that one instructor failed to assist the RL agent as she gave too many negative feedbacks.

In the second experiment [14] a human instructor assisted a humanoid robot in a sorting game. The goal was to give red balls to the instructor and to throw green balls away. The affective feedback

was detected by a facial expression reader [15]: smiling (positive) for confirmation of an action and frowning (negative) for correction (see Fig. 2 and 3). The robot successfully learned the desired task and was able to sort the last two balls without assistance. Furthermore, it proved that coaching by affective feedback leads to a significant improvement of HRI as the human instructor can act in a very natural (human-like) manner [14].

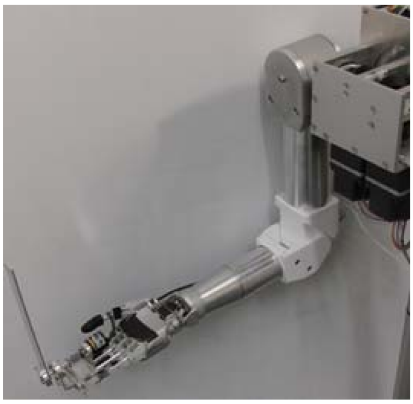


Figure 1. Robotic arm swinging up and keeping a pendulum balanced (figure taken from [17]).

Figure 2. Interaction with positive feedback (figure made available by Anna Gruebler).

3 THE METHOD OF LEVELS OF ABSTRACTION

Floridi proposes the method of levels of abstraction to analyze a system at different epistemic levels (cf. [10], ch. 3). This method to analyze all kinds of systems is inspired by the so-called Formal methods, a technique of computer science that aims at modeling a computer system regarding the "initial statement of a customer's requirements, through system design, implementation, testing, debugging, maintenance, verification, and evaluation" ([30], p. 8). Floridi utilizes this approach to evaluate a system technically for an epistemic analysis. In the line of Kant's critical philosophy [20], he stresses the epistemological issue to consider "the conditions of possibility of the

Figure 3. Interaction with negative feedback (figure made available by Anna Gruebler).

analysis (experience) of a particular system" ([10], p. 60). This recourse to the conditions of possibility of an analysis is crucial in order to avoid the mistake of analyzing a system independently of any specification of the analysis. These specifications, firstly, comprise the goal or purpose of an analysis. Furthermore, based on the general distinction that we can analyze a given system regarding its onotological levels of organization (LoO) and epistemological levels of explanation (LoE), levels of abstraction (LoA) serve to make explicit the ontological and epistemological commitments of the analysis. Thus, LoA guide the analysis teleologically towards a certain goal of interest.

As an epistemic levelism, each LoA depends on a certain observation or interpretation of a system. Hence, the technical concepts of the method of levels of abstraction and their formal definitions mainly comprise typed variables and observables, defined as follows (following quotations are from [10], ch. 3.2):

1. "A *typed variable* is a uniquely named conceptual entitiy (the variable) and a set, called its type, consisting of all the values that the entity may take." (p. 48)
2. "An *observable* is an interpreted typed variable, that is, a typed variable together with a statement of what feature of the system under consideration it represents." (ibid.)
3. "A *level of abstraction (LoA)* is a finite but non-empty set of observables." (p. 52)

Different LoAs of a system are integrated in a Gradient of abstraction (GoA), i.e. a LoA allows to specifically model a system, whereas in a GoA we can switch between different LoAs. To facilitate such a leveled analysis of a system, certain relations on a LoA and between all LoAs of a GoA must hold. The LoA-specific constraint is defined in terms of behavior:

4. "the *behaviour* of a system, at a given LoA, is defined to consist of a predicate whose free variables are observables at that LoA. The substitutions of values for observables that make the predicate true are called the system behaviours. A moderated LoA is defined to consist of an LoA together with a behaviour at that LoA." (p. 53)

Based on moderated LoAs, the GoA is defined as follows:

5. "A *gradient of abstractions*, GoA, is defined to consist of a finite set $\{L_i \mid 0 \leq i < n\}$ of moderated LoAs L_i, a family of relations

$R_{i,j} \subseteq L_i L_j$, for $0 \leq i \neq j < n$, relating the observables of each pair L_i and L_j of distinct LoAs in such a way that:

(a) the relationships are inverse: for $i \neq j$, $R_{i,j}$ is the reverse of $R_{j,i}$

(b) the behaviour p_j at L_j is at least as strong as the translated behaviour $P_{R_{i,j}}(p_i)$." (p. 55)

The GoA applied in our analysis of the coaching RL agent will be a *nested* GoA, i.e. its "non-empty relations are those between L_i and L_{i+1}, for each $0 \leq i < n-1$, and moreover the reverse of each $R_{i,i+1}$ is a surjective function from the observables of L_{i+1} to those of L_i." (p. 56)

Observations at one LoA can generally be related to observations at another LoA, but there are different ways of relating LoAs. Most prominently are hierarchical GoAs that propose one detailed LoA that serves to explain the observations at a more abstract LoA. This is for example the case in neurophysiological models of cognitive abilities where the biochemical reactions form the basic LoA. Cognitive abilities are modeled at more abstract or higher levels so that the observables at a higher level (e.g. phenomenal experience) can be translated to observables at a lower level (neurophysiological reactions). Whereas a hierarchical GoA can imply a reductionist approach, we make use of a net of abstractions, i.e. it is not our goal to reduce mental abilities to computational processes. Hence, we do not follow an ontological approach in order to determine the nature of mental or subjective states. Instead, we follow a *functional* approach in order to make explicit the functional organization of the coaching RL agent's information space [10], ch. 3.4.7. Accordingly, different LoAs are related by simulation, i.e. one LoA simulates the behavior of another LoA. The simulation relation connects different LoAs by a mapping relation R that relates the observables of two LoAs mutually. Unlike a hierarchical GoA or even a reductionist model of explanation, there is no basic or foundational LoA that realizes other LoAs unidirectionally. Instead, one system (here the coaching RL agent) is considered in terms of different functional realizations that are mutually related by a simulation relation. In a nested GoA, for every single observable at one LoA, it can be shown how this observable behaves at another LoA. In this way, different LoAs can be connected and serve as *mutual* explanation of their behavior. According to this mutual explanation of behavior the GoA serves to link different epistemic LoAs.

Our analysis of the coaching RL agent is placed in the broader context of subjective computing that was used to solve a learning task (see section 2.2 and 2.3). More precisely, we want to determine to what extent the algorithmic implementation can be related to a mental description of the agent's behavior. As mental abilities presuppose a subject that acts mentally, our analysis concerns the relational structure of the agent's information space. By means of this relational analysis, firstly, the kind of relations that hold between the agent and its environment (relation to others) and within the agent (self-referentiality) can be made explicit. By means of this relational account, we can, secondly, decode mentalistic terms (observing, considering, interpreting) in terms of the other LoAs and finally determine to what extent the coaching RL agent can be accounted for as exhibiting mental and therefore subjective abilities.

This way of analyzing subjective abilities of a robotic agent might force the straightforward objection that mental or subjective abilities are haphazardly imposed on a system that does not really possess these. This objection is grounded in the method of LoA as every LoA is based on an abstraction of the system under consideration: an abstraction of certain features is only possible if certain other features are neglected. E.g., we can analyze a robotic system regarding the mechanics of its components, the programming framework, the costs of its production, or, as in our case, its relational structure. Taking into consideration one perspective onto a system, implies neglecting other possible perspectives. Regarding the coaching RL agent, we neglect any phenomenal description of its behavior as we focus on the relational structure. Accordingly, we may not expect to analyze the agent's (potentially) mental behavior in human-like psychological terms. In the face of full-blown human subjectivity, it has to be admitted that the ascription of mental or subjective states cannot be completely justified by means of a relational GoA as its observables are defined regarding the system under consideration (here the coaching RL agent). To compare with a human subject we would have to define observables that also cover human cognition. But a GoA is always related to a certain system, and our goal is not to compare the coaching RL agent with a human subject (a futile undertaking in that the robot is without any doubt less subjectively and cognitively equipped), but to investigate certain relational aspects that are constitutive for subjectivity in general. If these relational aspects of cognition are utilized for the design of an agent and this agent shows a significantly improved and more intelligent behavior than without these subjective features, the technical implementation of certain relational aspects of subjectivity may be interpreted as a confirmation for the underlying philosophical concept of subjectivity. The methodological presupposition that justifies this ascription of mental abilities in favor of relational subjectivity, is based on a constructionist or design approach in philosophy [10], p. 72, 76ff.: a theoretical concept is validated and, at its best verified, if it is possible to design and implement a technical system according to this concept. Or, as in our case, if a technical implementation is shown to utilize this concept successfully.[5]

4 LEVELED ANALYSIS OF THE COACHING RL AGENT

Based on the method of LoAs we defined four LoAs in order to analyze the coaching RL agent:

1. *Algorithmic level.* This level depends on the algorithm that is implemented in the coaching RL agent. Whereas the computational level is fully covered by the original experiment [17], we focus on the cognitive abilities that are facilitated by the algorithm.
2. *Functional level.* The basic algorithm instantiates certain functions and therefore enables the agent to fulfill certain computational tasks; accordingly the agent determines, compares, and processes given feedback.
3. *Relational level.* The agent's information space depends on different kinds of relations to given input. For the following analysis it will be crucial to distinguish between a straightforward determination by direct world-coupling and a self-determination by means of a social relation that allocates feedback.
4. *Mentalistic level.* This level comprises the mentalistic description of the agent's actions. The goal of this analysis is to investigate to what extent the algorithmic, functional, and relational level allow for a mentalistic and therefore subjective characterization of the coaching RL agent.

Before we go into a formal treatment in order to bring forward a nested GoA of the coaching RL agent, we offer an informal treatment

[5] The constructionist approach asks for a continuative justification that exceeds the scope of this paper; see [10] for further discussion.

of coaching a robot. This serves to make clear at which levels we analyze the agent and how we relate the cognitive and interactional capacities to the mentalistic description of the agent's behavior. Based on this informal and the subsequent formal treatment it will be possible to evaluate to what extent the ascription of mental abilities is justified.

4.1 Instantiating a subjective agent in social interaction

The task of coaching a robot offers an instructive way to study the behavior of an autonomous agent that interacts with humans. One special feature consists in the mutual exchange between the robotic agent and the human trainer. The agent is not only supposed to deliver a computational result as for instance in the case of search engines, but its actions provoke the trainer's feedback that itself serves the agent to modify its actions. Even if the exchange between robot and human does not take place on a linguistic level, the trainer's feedback is answered by the robot's behavior whereas the behavior provokes new feedback. To improve the learning abilities of the robotic agent a RL agent was complemented with a coaching function (see section 2.2). This functional follows two central purposes: By means of the feedback the RL agent can adjust the learning parameter (reward function) that defines the success of an action during the learning process. On the other hand the coaching function enables a human trainer to interact with a robotic agent in a very natural (i.e. affective) manner. The trainer just gives positive or negative feedback that is to be processed autonomously (interpreted) by the robotic agent. By allocating the feedback by means of a facial emotion reader [15] the mental workload for the human trainer decreases to a minimum level that does not differ significantly from a human-human interaction.

Our case study [17] is based on a robot arm platform (see Fig. 1). The robot has to solve the task of keeping a pendulum balanced. In order to accomplish this task the agent can modify the joints of its arm to handle the pendulum. But it has to learn how to modify its joints. In the coaching framework a human trainer gives a two digit feedback (positive or negative) while the agent is trying to keep the pendulum balanced. Accordingly, the robot must be able to process the feedback. The final goal is that the agent processes the feedback and adjusts its actions autonomously. As we deal with a robotic system we basically have to take into account the algorithmic implementation of the cognitive abilities required to process the feedback. So the basic LoA comprises of the algorithmic implementation.[6] Accordingly, at this level we should not conceive of an agent that acts, but of algorithmic processing. In our case study the robotic agent is able to react to feedback in a twofold manner. The algorithm enables the robot to determine which of its behavior refers to a feedback. This step of determining the feedback's target behavior (causality detection) is crucial for the processing of the feedback as the robot must be able to relate a feedback to its behavior. Even in human-human learning we know the common misunderstanding that the trainee sometimes allocates the feedback to a different behavior as the trainer aimed at. Furthermore, the algorithm allows the coaching RL agent to compare a feedback with previous feedbacks related to the same action. This test for consistency serves to identify contradictory feedback as an action cannot be conceived of as a successful action based on positive feedback when at the same time the action was evaluated negatively earlier. Again, the consistency of feedback is even crucial for human-human learning as a trainee can benefit from unambiguous feedback whereas contradictory feedback already presupposes a certain level of expertise if the trainee is supposed to profit in the same way as in the case of unambiguous feedback. Finally, when a feedback was assigned to a certain target-behavior and the feedback is consistent with previous feedbacks of this behavior, then the algorithm leads to a modification of the reward function and subsequently to adapted behavior. This final adaption of behavior can count as a successful learning process as the robot's behavior improved in order to accomplish the task to a higher degree than before the learning process.

In our example the learning process goes like this: When trying to balance the pendulum, the robotic arm platform starts with the initial posture of the pendulum as vertically downward. The robot decides how many degrees it moves its joint at every time step according to the current situation. Furthermore, it remembers the history of its actions. While balancing the pendulum, a human trainer gives positive or negative feedback. Via an interface this feedback is allocated as a reward value for the RL agent. Then every single feedback is processed according to the algorithm, i.e. the robot, firstly, determines the target behavior of a feedback. The target behavior of the feedback is the movement of the robot's joints within a certain time range. Hence, when the feedback is given from the trainer, the robot is able to estimate which of its actions the trainer actually evaluated by referring to a certain time range of the history of its actions. Whereas this time range, which the feedback refers to, can in principal also be learnt. In the experiment the time range was defined based on the measurement of human's delay of cognition. How much time passes before a human trainer gives feedback was measured: the results show that the minimum and maximum delay lies within 300 to 800[ms] (cf. [17], p. 11). According to this data, the coaching RL agent mapped a feedback to its behavior 300 to 800[ms] ago. After the determination of target behavior the agent, secondly, compares the feedback to previous feedbacks of the same behavior. If the previous acceptance or denial of this behavior is confirmed, the robot modifies its reward function accordingly. Based on this adjusted reward function the agent prefers the actions that were evaluated positively and changes its action rules. Thus, when the feedback confirmed a certain modification of the joints, then the agent will modify its behavior in order to move its joints according to the confirmed behavior. If, for example, the position of one joint within a certain range provoked positive feedback, then the robot will not exceed this range. Or if a certain joint angle provoked only negative feedback, the robot will not move this joint any more to this degree.[7]

Obviously, the previous description of the algorithmic level does not capture a mental or subjective ability. It entails the description of data processing and the transformation of data into modified behavior. But when we conceive this algorithmic processing at a functional level, we can take into account the functions instantiated by the algorithm. The functional description refers to the causal role of a component and specifies how an agent is empowered to act [4], [3]. The functional level allows to abstract from the algorithmic as computational processes and conceive the latter as cognitive functions of an agent. This shift of our investigation is crucial as on the algorithmic level there is strictly speaking no agent acting, but an algorithm is processing data. The fact that the algorithm enables an agent cannot be made explicit until we shift our attention to a functional level. Here it is that the computational reward value becomes a feedback as

[6] The study of a robotic system, or more generally, of an algorithm guarantuees that the system is controllable and implementable, i.e. we deal with a *white box* so that all parameters and internal operations can clearly be specified (cf. [13]).

[7] In the actual experiment, the ability of interpretation was limited to the extent that the robot could not process prevailing negative feedback.

a feedback is only possible in the mutual exchange of agents, i.e. between the human trainer and the robotic agent. The trainer primarily interacts with the robotic agent and not with the algorithm. Whereas in a strictly computational perspective one might say that the human trainer interacts with the algorithm, this does not make sense if we investigate the coaching process from a cognitive perspective. Cognitively speaking the computational reward is a feedback that has to be translated into a computational format. But again, the human trainer is not directly giving a computational reward value but an affective feedback [15]. Thus, the whole importance of the difference between function and algorithm lies in the transformation of an affective reaction (positive/smile or negative/frown) into a binary reward value. Or, correspondingly, i.e. seen from algorithm to function, in the empowerment of an agent to operate on affective feedback. Hence, in a functional perspective we can actually conceive of a robotic *agent* that receives feedback. Functionally speaking, it is an agent that determines target behavior, compares and finally processes feedback. We shifted from an algorithmic description of computational processes to a functional characterization of an agent.

Whereas we proceeded from algorithmic processing to the capacities of an agent, the functional characterization still does not allow for a mental or subjective description of the coaching RL agent. Certain functions can be instantiated by many different systems that are obviously far from being mental or subjective. A thermostat fulfills the function of adjusting temperature or a search engine ranks data according to some specified criteria. So we have to take into account a further LoA that helps to identify if and to what extent the coaching RL agent is supposed to act subjectively. This is the relational LoA that models the agent's relations to itself and others. When conceiving of mental abilities (implied in the use of mentalistic language), we expect an agent that acts autonomously and is not just responding automatically to some input data. Hence, the agent's relations to some given input is crucial for evaluating its behavior [16].

Based on a relational analysis we can distinguish between different kinds of relations between the agent and its environment. On the one hand the agent's behavior is forced by standard RL that is based on direct world coupling. In standard RL, the behavior gets automatically modified by environmental constraints. This modification depends on the reward function as the criteria which actions count as a success and which actions fail to accomplish the task. In fact, our example displays an extreme case as when the pendulum fell down no further adjustment or modification of behavior is possible. The task inevitably failed. But in more flexible tasks, e.g. as in the case of navigation, environmental constraints could force an agent to change its direction when it encounters an obstacle. The crucial point here is that the agent's relation to an input (the obstacle) is determined, i.e. the agent's behavior changes automatically without that the agent does have any control of this modification of its behavior. Furthermore, all modifications depend on the predefined reward function. In the case of the coaching RL agent, this way of direct world-coupling is complemented by an autonomous self-relation. The robotic arm not only reacts automatically to external events (here that the pendulum falls down). The agent is able to operate on the automatic learning process so that this process is not any more comletely determining the agent's behavior. Based on the algorithmic causality and error detection, or the functional capacity to determine target behavior and compare feedback respectively, the agent is able to process a binary feedback and decide by itself whether and to what extent its behavior should be modified. Both relations, the direct world-coupling and the interpretation of feedback, contribute to the agent's performance.

One might object that the robotic arm does not engage in full-blown decision making, but that is not the point here. Here it is crucial that the agent's behavior is significantly improved in that the final modification of the behavior is left to the agent itself. The agent operates autonomously on feedback and therefore relates autonomously to its own internal model of the trainer's evaluation. Thus, autonomous self-referentiality comprises that an agent operates on its own internal states whereas these operations do not completely underlie any external constraints [16]. The underlying concept of autonomy does not aim at complete self-determined behavior. Instead, autonomous behavior can be generated in opposition to determined behavior, i.e. the determination of the agent gets limited, or, correspndingly, the agent's autonomous capacity has to be constrained in order to bring forward successful behavior. The theorem of 'realization by means of restriction'[8] clarifies the role of social interaction. In our case, social interaction lies between the autonomous interpretation and direct world-coupling, i.e. it is a *partial* determination of the agent's information space as the agent is constrained by the feedback values, but is autonomous regarding their further processing. Due to the difficulties to define a suitable reward function a priori (see section 2.2), the coaching function and the feedback were introduced in order to assist the robot with updating the reward function. Accordingly, the subjective momentum of the agent's informational space depends on a mutual dependence of the autonomous self-relation and the social interaction: in order to evaluate its behavior autonomously the agent depends on a certain input (feedback) that confines his capacity to interpret the feedback to some reasonable options. Otherwise, the agent would have no criteria how to evaluate its actions, i.e. how to move its joints.

The autonomous and at the same time partially determined behavior lies at the ground of a subjective agent and serves to identify the final LoA. The coaching RL agent can be regarded as acting mentally in that it interprets the feedback based on an autonomous decision making: the agent *considers* contingency, *observes* the consistency of given feedback which results in an interpretation. Mental states of considering, observing, and interpreting that presuppose a subjective agent are based on the mutual relationship of autonomous self-referentiality and social interaction in that the straightforward determination of behavior by direct world-coupling is interrupted. We can call the agent's interpretation 'mental' or 'subjectiv' as this behavior is finally determined in the agent's information space by the agent itself and not primarily by some external constraints. The robot, being socially receptive for direct interaction with a human and its autonomous decision making, qualifies the coaching RL agent as a basically subjective agent. Again, one might object that this kind of subjectivity is less than what we usually ascribe to full-blown human subjects. But despite these obvious restrictions, the leveled analysis of the robotic agent offers us an account of subjectivity that does not rely on intractable phenomenal or psycgological states. We can instead follow the generation of a subjective agent from scratch. Furthermore, we are forced to include social interaction, which easily gets lost in phenomenal accounts. The main purpose of the following formal treatment lies in the need to make explicit the relations within and between every LoA, as subjectivity is here primarily seen under a relational viewpoint.

4.2 Nested GoA of the coaching RL agent

According to the previous stated method of levels of abstraction and the informal treatment, the RL agent is now to be analyzed formally

[8] See the chapter on schematism in [20], and [10].

at four LoAs. Each LoA comprises three observables (interface, interpretation, learning) with specific variables related to the observables. The relational LoA forms an exception, in that not the observables themselves but the relational structure of the agent's processing describes the behavior of this LoA. The following formalization does not depend on any specific mathematical standard but merely seeks to make clear the different levels of the agent's cognitive activity and especially the relations between the agent and the trainer's feedback at L_2.

The nested GoA is based on the following levels (L), comprising the observables interface, interpretation, learning, and corresponding variables:

- L_0: *algorithmic level*
  - Interface: reward value V
  - Interpretation: estimation of reward function E_F
  - Learning: updating reward function and action rules U
- L_1: functional level
  - Interface: feedback F
  - Interpretation: estimation of relevance E_R
  - Learning: processing feedback F_p
- L_2: relational level
  - Agent's self-referentiality: A_s
  - Agent's social relation (interaction): A_i
  - Direct world coupling (standard RL): A_d
- L_3: mentalistic level
  - Interface: social receptivity S
  - Interpretation: interpreting feedback F_i
  - Learning: reflecting feedback F_r

These observables and corresponding variables form a nested GoA of the coaching RL agent (see table 1). The GoA consists of four LoAs specified in the first column and beginning with the algorithmic level. Due to the epistemic foundation of LoAs, the epistemic regard according to which the coaching RL agent is interpreted is given at each level. Each LoA consists of three observables: interface, interpretation, and learning. On L_0 the system is analyzed regarding its algorithmic processing. This computational level is fully covered by the original experiment [17]. In the following analysis we therefore solely focus on those aspects concerning the instantiation of an information space which is computationally implemented as a continuous state-action-space: a certain signal, the reward value V is delivered by an interface and gets processed in the course of causality and error detection. In the course of interpretation, these processes of detection are regarded as an estimation of the reward function E_F. Learning at the algorithmic level consists of an updated reward function and correspondingly updated action rules U. The system's behavior can be specified by the use of the following predicates: V delivers a binary value corresponding to a positive (smile, V^+) or negative (frown, V^-) evaluation of the instructor. E_F delivers a value under or above the current reward function and leads in the first case to an update U of the reward function and the action rules so that U contains the updated and modified reward function that will result in adapted behavior.

At the subsequent LoAs these processes remain the same, but are analyzed differently. Considering the functionality of the algorithm at L_1, the algorithm enables an agent to fulfill certain computational tasks: the agent determines, compares, and processes given feedback. This functional mapping serves to identify the cognitive processes of the agent (cf. section 4.1) as follows: The algorithmic observables at L_0 are mapped to L_1 as follows: V functions as feedback and takes the values of 'confirmation' V^+ or 'correction' V^-, i.e. the function of the computational values consists in allocating positive or negative feedback. Hence, the meaning of the computational value V for the agent's behavior is identified by its function. The same counts for the estimation of the reward function E_R: E_F fulfills the cognitive function of specifying the feedback according its relevancy for the agent's behavior. E_R is the result of estimating V, i.e. $E_R = E_F(V)$. Finally the cognitive function of U is processing feedback by updating the reward function and the action rules in order to increase learnability, i.e. $F_p = U(E_F)$.

L_2 contains the crucial relational analysis. The agent's information space depends on two kinds of relations that hold between the previous observables and the agent's capacity to operate on them. On the one hand the agent underlies two determining relations to others: Based on the implementation of standard RL the coaching RL agent depends on external and automatic determination of behavior through direct world coupling A_d and a subsequent adaption to the environment. Secondly, in the course of social interaction A_i the trainer allocates feedback F. Whereas the feedback contains a fixed value (positive or negative) that cannot be altered by the agent, the further processing of the feedback is subject to an interpretation by the coaching RL agent that decides if its behavior gets modified. Hence the degree of determination of the agent's behavior decreases significantly in the course of social interaction. The subjective momentum of the agent's information space is generated by the second kind of relation, i.e. the agent's autonomous relation to its own internal model of the trainer's evaluation. The RL agent is able to modify the reward function and action rules autonomously and therefore indirectly its behavior. This relation to the incoming feedback is autonomous as the latter does not determine the agent necessarily or immediately as is the case with standard RL. Opposed to externally determined behavior as the result of A_d, subjectively modified behavior is instantiated by autonomous acts of interpretation by the agent itself. Thus, the agent's subjective information space depends on these simultaneous relations as can be made explicit by mapping the observables at L_1 onto L_2: According to standard RL the agent is determined directly through direct world coupling A_d. At the same time feedback is allocated by social interaction $A_i[F]$ that constrains the autonomous modification of the reward function: F is processed by E_R to F_p depending on the agent's own, i.e. subjective, interpretation of the feedback $A_s[E_R(F) \rightarrow F_p]$. The agent's autonomy consists in its ability to modify its own learning process by adjusting the reward function by itself. From a relational viewpoint, the agent's subjective determination of the reward function is constituted simultaneously with an objective determination of its behavior by direct world-coupling (see section 4.1).

The complete behavior of the coaching RL agent at L_2 depends on these parallel processes. Whereas determined behavior alone is not a special characteristic of *subjective behavior*, autonomous self-referentiality (A_s) and social interaction (A_i) are relevant for the final LoA, the mentalistic level. The subjective ability to modify the automatic learning process by autonomously processing feedback forms a necessary condition for subjective computing. At the same time autonomous self-referential behavior can only be effectively utilized in the course of social interaction as the agent has to learn how to modify its learning process. Hence, autonomous self-

Table 1. Nested GoA of the coaching RL agent.

LoA *Relations*	**Observables**		
L_0: algorithmic (algorithmic processing)	**Interface** reward value V	**Interpretation** causality detection $\rightarrow$ error detection, i.e. estimation of the reward function E_F	**Learning** updating reward function $\rightarrow$ updating action rules U
$R_{0,1}$: mapping algorithm to functions	$R_{0,1}(V, F)$ $F = V^+ \vee V^-$	$R_{0,1}[E_F, E_R]$ $E_R = E_F(V)$	$R_{0,1}(U, F_p)$ $F_p = U(E_F)$
L_1: functional (functions realized by the algorithm)	**Interface** feedback F	**Interpretation** agent determines target behavior (contingency) $\rightarrow$ compares feedbacks (consistency), i.e. specifies feedback by estimating its relevance E_R	**Learning** agent processes feedback F_p
$R_{1,2}$: mapping functions to relations	$R_{1,2}[(F, E_R, F_p), (A_s, A_i), A_d]$ $A_s[E_R(F) \rightarrow F_p] \wedge A_i[F] \wedge A_d$		
L_2: relational (relational structure of agent's processing)	**Self-referentiality** agent relates (based on autonomous acts of estimating) to its own internal model of the trainer's evaluation, i.e. an autonomous and subjective self-relation A_s	**Relations to other** *Social relation* trainer's evaluation of the agent's behavior (feedback F) is allocated in social interaction A_i	 *Direct world coupling* determined and objective relation A_d based on direct world coupling (standard RL)
$R_{2,3}$: mapping relations to mental abilities	$R_{2,3}(A_s, S)$ $S = A_s(F)$	$R_{2,3}(A_s, F_i)$ $F_i = A_s(E_R)$	$R_{2,3}(A_s, F_r)$ $F_r = A_s(F_p)$
L_3: mentalistic (mental abilities)	**Interface** social receptivity S	**Interpretation** agent considers contingency, carefully observes consistency of given feedback, i.e. interprets feedback F_i	**Learning** agent learns autonomously, i.e. agent reflects feedback or differences of coaching strategies by its behavior F_r

referentiality and social interaction interdependently enable a subjective agent. Again, subjectivity here means that the robotic agent is able to modify an ongoing automatic process whereas this modification is externally supported (here by feedback) but is finally left to the agent's decision. Those subjective and interactional issues arise in scenarios where a robotic agent is supposed to adopt a task and to accomplish this task autonomously (e.g. driving assistance, search and rescue applications, or autonomous control in hybrid assistive limb [**?**]). But due to the difficulties of defining the robot's actions in advance or to define a suitable reward function a priori, social interaction (coaching) can be utilized in order to support the robot's autonomous modification of its behavior and therefore improve its learnability.

The relational structure and the instantiation of a subjective relation in the agent's information space finally allow for a mentalistic interpretation of the coaching RL agent at L_3. Usually, we ascribe acts like considering and reflecting to a full-blown subject. This is, obviously, not the case here. Full-blown subjectivity depends on further features like natural language and ethical addressability. But when taking into account the social interaction of the coaching RL agent, this agent acts as an autonomous counterpart of the human, i.e. the agent exhibits a sufficient level of autonomy that we can ascribe mental activity to it as follows: in operating on the instructor's input, i.e. autonomously relating to the feedback $A_s(F)$, the agent becomes so-

cially receptive S in the course of interaction. The RL agent shows subjective behavior when individually and situation-dependently interpreting feedback $F_i = A_s(E_R)$ and correspondingly learning by updating the reward function and action rules according to its interpretation, i.e. autonomously processing or reflecting feedback $F_r = A_s(F_p)$. The social interaction between the trainer and the robotic agent is crucial for A_i and the mentalistic character of the RL agent's behavior as the feedback offers an additional input (binary cues) opposed to strict world-coupling in standard RL. The subjective momentum, based on autonomous self-referentiality, occurs as the RL agent's non-deterministic consideration of contingency and observation of consistency of feedback as well as in the subsequent reflection of differences of coaching strategies by means of more or less successful learning. There is no predefined reaction or development of the coaching RL agent's behavior, but subjective behavior due to the internal indeterminacy of the modification of the learning process. At the same time the agent's autonomous ability relies on social interaction that guides its ability to modify its learning process. Without this guidance the agent would not be able to execute its autonomous modification of the reward function as it has no information how and to what extent a modification might support to accomplish its task.

5 CONCLUSION

We wanted to investigate what it can mean for a robotic agent to behave subjectively. We approached this question by analyzing to what extent mental abilities can be ascribed to a robotic agent. In the course of analyzing a coaching RL agent at four LoAs we made explicit a relational level (L_2) that shows how mental abilities can be ascribed to the agent: the coaching RL agent behaves subjectively in that it is able to modify its own automatic learning processes by means of feedback that is allocated in social interaction. At the same time, the agent is still being determined by direct world-coupling. Hence, the relational level confirms a relational notion of subjectivity.

On the other hand, this result underlies a certain caveat in that the nested GoA of the coaching RL agent is based on an abstraction that focuses on the relational structure of the agent, i.e. we analyzed to what extent the agent's actions are self-referential and related to others as well as self-determined and externally determined. This relational account of the robot's information space does not cover a common psychologically or phenomenally based description of human-like cognitive processes as it is mainly decoupled from the concept of consciousness and linked to intelligence. From a relational viewpoint, consciousness is regarded as cognitive product. Hence, it is necessary to go back to a level of abstraction that does not presuppose any conscious states if conscious, or less difficult, mental abilities have to come into reach of an explanation. By modeling relational features of intelligence by means of a technical implementation, we gained an analysis of cognitive abilities that is fully tractable and implementable.

Based on a technical implementation that showed a significant improvement of an agent's behavior by means of the coaching function, it was relationally justified to conclude that the coaching RL agent acts subjectively as it makes effective use of autonomous self-referentiality and social interaction. The agent's subjectivity is generated in this course of action as the agent's self-determined behavior opposed to external determination by direct world-coupling. By means of this relational abstraction of the coaching RL agent, we can link the technical implementation with the conceptual foundation of subjectivity and subjective computing, respectively. With regard to the further development of subjective agents, the link of the technical and theoretical domain supports the improvement of subjective abilities. The theoretical framework of relational subjectivity can guide an extension of self-referential processing in order to allow the coaching RL agent to process ambiguous feedback. Another open question concerns social interaction in other modes than binary feedback. With regard to full-blown human subjectivity, the relational account does not exclude modeling more complex cognitive abilities as the use of natural language or ethical addressability. On the other hand, the theoretical framework of relational subjectivity is being modeled in the course of technical implementation. This allows us to test and verify a relational modeling of subjectivity.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] Cynthia Breazeal, 'Toward sociable robots', *Robotics and Autonomous Systems*, **42**(3-4), 167–175, (March 2003).
[2] Daniel Breazeale and Günter Zöller, *The system of ethics: according to the principles of the Wissenschaftslehre*, Cambridge University Press, Cambridge, UK; New York, 2005.
[3] Mark B. Couch, 'Causal role theories of functional explanation', *Internet Encyclopedia of Philosophy*, (2011).
[4] Robert Cummins, *The Nature of Psychological Explanation.*, MIT Press, Cambridge, Mass., 1983.
[5] Kerstin Dautenhahn, 'A paradigm shift in artificial intelligence: Why social intelligence matters in the design and development of robots with Human-Like intelligence', in *50 Years of Artificial Intelligence*, eds., Max Lungarella, Fumiya Iida, Josh Bongard, and Rolf Pfeifer, volume 4850 of *Lecture Notes in Computer Science*, 288–302, Springer Berlin/Heidelberg, (2007).
[6] Kerstin Dautenhahn, Alan H. Bond, Lola Canamero, and Bruce Edmonds, *Socially intelligent agents: creating relationships with computers and robots*, volume 3 of *Multiagent Systems, Artificial Societies, and Simulated Organizations*, Springer, Berlin/Heidelberg, May 2002.
[7] Kerstin Dautenhahn, Bernard Ogden, and Tom Quick, 'From embodied to socially embedded agents – implications for interaction-aware robots', *Cognitive Systems Research*, **3**(3), 397–428, (September 2002).
[8] Johann Gottlieb Fichte, 'Grundlage der gesamten wissenschaftslehre als handschrift für seine zuhörer (1794)', in *Fichte-Gesamtausgabe der Bayerischen Akademie der Wissenschaften. Bd. I, 2.*, eds., Reinhard Lauth and Hans Jacob, 251–451, Bad Cannstatt, (1962).
[9] Johann Gottlieb Fichte, 'Wissenschaftslehre 1811', in *Fichte-Gesamtausgabe der Bayerischen Akademie der Wissenschaften. Bd. II, 12.*, eds., Reinhard Lauth and Hans Jacob, 138–299, Bad Cannstatt, (1962).
[10] Luciano Floridi, *The Philosophy of Information*, Oxford University Press, Oxford, 2011.
[11] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn, 'A survey of socially interactive robots', *Robotics and Autonomous Systems*, **42**(3-4), 143–166, (2003).
[12] Michael A. Goodrich and Alan C. Schultz, 'Human-robot interaction: a survey', *Found. Trends Hum.-Comput. Interact.*, **1**(3), 203–275, (January 2007).
[13] Gian Maria Greco, Gianluca Paronitti, Matteo Turilli, and Luciano Floridi, 'How to do philosophy informationally', in *Lecture Notes on Artificial Intelligence*, volume 3782, pp. 623–634, (2005).
[14] Anna Gruebler, Vincent Berenz, and Kenji Suzuki, 'Coaching robot behavior using continuous physiological affective feedback', in *2011 11th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pp. 466–471. IEEE, (October 2011).
[15] Anna Gruebler and Kenji Suzuki, 'Measurement of distal EMG signals using a wearable device for reading facial expressions', in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4594–4597. IEEE, (September 2010).

[16] Patrick Grüneberg, *Projektives Bewusstsein. Th. Metzingers Selbstmodelltheorie und J.G. Fichtes Wissenschaftslehre*, Dissertation, Technische Universität Berlin, 2012.
[17] Masakazu Hirokawa and Kenji Suzuki, 'Coaching to enhance the online behavior learning of a robotic agent', in *Knowledge-Based and Intelligent Information and Engineering Systems*, eds., Rossitza Setchi, Ivan Jordanov, Robert J. Howlett, and Lakhmi C. Jain, volume 6276, 148–157, Springer, Berlin/Heidelberg, (2010).
[18] Axel Honneth, *The struggle for recognition: the moral grammar of social conflicts*, MIT Press, Cambridge, MA, 1996.
[19] Frank Jackson, 'Epiphenomenal qualia', *The Philosophical Quarterly*, **32**(127), 127–136, (April 1982).
[20] Immanuel Kant, *Kritik der reinen Vernunft (1781/1787)*., Hamburg, 1990.
[21] Norimasa Kobori, Kenji Suzuki, Pitoyo Hartono, and Shuji Hashimoto, 'Reinforcement learning with temperature distribution based on likelihood function', *Transactions of the Japanese Society for Artificial Intelligence*, **20**, 297–305, (2005).
[22] Thomas Metzinger, *Being no one. The Self-model Theory of Subjectivity.*, MIT Press, Cambridge, Mass., 2003.
[23] Anthony F. Morse, Carlos Herrera, Robert Clowes, Alberto Montebelli, and Tom Ziemke, 'The role of robotic modelling in cognitive science', *New Ideas in Psychology*, **29**(3), 312–324, (2011).
[24] Thomas Nagel, 'What is it like to be a bat?', *Philosophical Review*, **83**(4), 435–450, (1974).
[25] Momoko Nakatani, Kenji Suzuki, and Shuji Hashimoto, 'Subjective-evaluation oriented teaching scheme for a biped humanoid robot', *IEEE-RAS International Conference on Humanoid Robots*, (2003).
[26] Marcia Riley, Ales Ude, Christopher Atkeson, and Gordon Cheng, 'Coaching: An approach to efficiently and intuitively create humanoid robot behaviors', in *2006 6th IEEE-RAS International Conference on Humanoid Robots*, pp. 567–574. IEEE, (December 2006).
[27] Richard S. Sutton and Andrew G. Barto, *Reinforcement learning: an introduction*, MIT Press, Cambridge, Mass., 1998.
[28] Andrea L. Thomaz and Cynthia Breazeal, 'Teachable robots: Understanding human teaching behavior to build more effective robot learners', *Artificial Intelligence*, **172**(6-7), 716–737, (April 2008).
[29] Andrea Lockerd Thomaz. Socially guided machine learning. http://dspace.mit.edu/handle/1721.1/36160, 2006. Thesis (Ph. D.)–Massachusetts Institute of Technology, School of Architecture and Planning, Program in Media Arts and Sciences, 2006.
[30] Jeannette M. Wing, 'A specifier's introduction to formal methods', *IEEE Computer*, **23**(9), 8–24, (September 1990).

Bill Gates is not a Parking Meter: Philosophical Quality Control in Automated Ontology-building

Catherine Legg and **Samuel Sarjant**[1]

Abstract. The somewhat old-fashioned concept of philosophical categories is revived and put to work in automated ontology building. We describe a project harvesting knowledge from Wikipedia's category network in which the principled ontological structure of Cyc was leveraged to furnish an extra layer of accuracy-checking over and above more usual corrections which draw on automated measures of semantic relatedness.

1 PHILOSOPHICAL CATEGORIES

`S1: The number 8 is a very red number.`

There is something clearly wrong with this statement, which seems to make it somehow 'worse than false.' For a false statement can be negated to produce a truth, but

`S2: The number 8 is not a very red number.`

doesn't seem right either.[2] The problem seems to be that numbers are not the kind of thing that can have colours — if someone thinks so then they don't understand *what kinds of things numbers are*.[3]

The traditional philosophical term for what is wrong is that `S1` commits a *category mistake*. It mixes kinds of thing nonsensically. A traditional task of philosophy was to identify the most basic categories into which our knowledge of reality should be divided, and thereby produce principles for avoiding such statements. One of the first categorical systems was produced by Aristotle, who divided predicates into ten groups (*Substance*, *Quantity*, *Quality*, *Relation*, *Place*, *Time*, *Posture*, *State*, *Action*, and *Passion*). The differences between these predicates were assumed to reflect differences in the ontological natures of their arguments. For example, the kinds of things that are earlier and later (*Time*) are not the kinds of things that are heavy or light (*Substance*). Category lists were also produced by Kant, Peirce, and many other Western philosophers.

We believe there is a subtle but important distinction between philosophical *categories* and mere *properties*. Although both divide entities into groups, and may be represented by classes, categories arguably provide a deeper, more sortal division which enforces *constraints*, which distinctions between properties do not always do. So for instance, while we know that the same thing cannot be both a colour and a number, the same cannot be said for green and square. However, at what 'level' of an ontology categorical divisions give way to mere property divisions is frequently unclear and contested. This has led to skepticism about the worth of philosophical categories which will now be touched on.

This task of mapping out categories largely disappeared from philosophy in the twentieth century.[4] The logical positivists identified such investigations with the "speculative metaphysics" which they sought to quash, believing that the only meaningful questions could be settled by empirical observation [4, 19].

Following this, Quine presented his famous logical criterion of ontological commitment: "to be is to be the value of a bound variable. . . [in our best scientific theory]" [22]. This widely admired pronouncement may be understood as flattening all philosophical categories into one 'mode of being'. Just as there is just one existential quantifier in first-order logic, Quine claimed, ontologically speaking there is just one kind of existence, with binary values (does and does not exist). Thus there are no *degrees* of existence, nor are there *kinds* — rather there are different *kinds of objects* which all have the same kind of existence.

This move to a single mode of being might be thought to reopen the original problem of why certain properties are instantiated by certain kinds of objects and not others, and why statements such as `S1` seem worse than false. A popular response — common in the analytic tradition as a reply to many problems — has been to fall back on faith in an ideal language, such as modern scientific terminology (perhaps positions of atoms and molecules), which is fantasized as 'category-free.'

Be that as it may, we will now examine a computer science research project which recapitulated much of the last 3000 years of philosophical metaphysics in a fascinating way.

2 THE CYC PROJECT

2.1 Goals and basic structure

When the field of Artificial Intelligence struggled in the early 80s with brittle reasoning and inability to understand natural language, the Cyc project was conceived as a way of blasting through these blocks by *codifying common sense*. It sought to represent in a giant knowledge base, "the millions of everyday terms, concepts, facts, and rules of thumb that comprise human consensus reality", sometimes expressed as everything a six-year-old knows that allows her to understand natural language and start learning independently [8, 9].

This ambitious project has lasted over 25 years, producing a taxonomic structure purporting to cover all conceivable human knowl-

[1] The University of Waikato, New Zealand, email: {clegg, sjs31}@waikato.ac.nz

[2] Some philosophers do take a hard line on statements such as `S2`, claiming that it is literally true, but it does at least seem to have misleading pragmatic implications.

[3] There is the phenomenon of synaesthesia. But the rare individuals capable of this feat do not seem to converge on any objective colour-number correlation.

[4] Notable exceptions: [5, 10, 11, 23].

edge. It includes over 600,000 categories, and over two million axioms, a purpose-built inference engine, and a natural language interface. All knowledge is represented in *CycL*, which has the expressivity of higher-order logic — allowing assertions about assertions, context logic (Cyc contains 6000 "Microtheories"), and some modal statements.

The initial plan was to bring the system as quickly as possible to a point where it could begin to learn on its own, for instance by reading the newspaper [8, 9]. Doug Lenat estimated in 1986 that this would take five years (350 person-years) of effort and 250,000 rules, but it has still not happened, leading to widespread scepticism about the project.

2.2 Categories and common sense knowledge

Nevertheless, it is worth noting that the Cyc project *did* meet some of its goals. Consider the following, chosen at random as a truth no-one would bother to teach a child, but which by the age of six she would know by common-sense:

`S3: Bill Gates is not a parking meter.`[5]

This statement has never been asserted into Cyc. Nevertheless Cyc knows it, and can justify it as shown in Figure 1.

```
BillGates is known not to be an instance of
      ParkingMeter in mt WikipediaToCycDataMt.
sbhl conflict: (isa BillGates ParkingMeter) TRUE
            WikipediaToCycDataMt
      because: (isa BillGates MaleHuman)
            True-JustificationTruth
(genls MaleHuman MaleAnimal) TRUE
(genls MaleAnimal Animal) TRUE
(genls Animal AnimalBLO) TRUE
(genls AnimalBLO BiologicalLivingObject) TRUE
(disjointWith BiologicalLivingObject
      Artifact-Generic) TRUE
(genls Technology-Artifact Artifact-Generic) TRUE
(genls MechanicalDevice Technology-Artifact) TRUE
(genls ParkingMeter MechanicalDevice) TRUE
```

Figure 1. Justification produced in ResearchCyc 1.0, 2009

The crucial premise is the claim of disjointness between the classes of living things and artifacts. The Cyc system only contains several thousand explicit `disjointWith`[6] statements, but as seen above, these ramify through the knowledge hierarchy in a powerful, open-ended way.

A related feature of Cyc's common-sense knowledge is its so-called *semantic argument constraints on relations*. For example `(arg1Isa birthDate Animal)` represents that only animals have birthdays. These features of Cyc are a form of categorical knowledge. Although some of the categories invoked might seem relatively specific and trivial compared to Aristotle's, logically the constraining process is the same.

[5] Presenting this material to research seminars it has been pointed out that there is a metaphorical yet highly meaningful sense in which Bill Gates (if not personally, then in his capacity as company director) does serve as a parking meter for the community of computer users. Nevertheless, in the kinds of applications discussed in this paper we must alas confine ourselves to literal truth, which is challenging enough to represent.

[6] Terms taken from the CycL language are represented in `TrueType` throughout the paper.

In the early days of Cyc, knowledge engineers laboured to input common-sense knowledge in the form of rules (e.g. "If people do something for recreation that puts them at risk of bodily harm, then they are adventurous"). Reasoning over such rules required inferencing of such complexity that they almost never 'fired' (were recognized as relevant), or if they did fire they positively hampered query resolution (i.e. finding the answer). By contrast Cyc's disjointness and semantic predicate-argument constraints were simple and effective, so much so that they were enforced at the knowledge-entry level. Thus returning again to `S1`, this statement could not be asserted into Cyc because redness is represented as the class of red things which generalizes to spatiotemporally located things, while numbers generalizes to abstract objects, and once again these high level classes are known to be disjoint in Cyc.

We believe these constraints constitute an untapped resource for a distinctively ontological quality control for automated knowledge integration. Below we show how we put them to work in a practical project.

3 "SEMANTIC RELATEDNESS"

When 'good-old fashioned' rule-based AI systems such as Cyc apparently failed to render computers capable of understanding the meaning of natural language, AI researchers turned to more brute, statistical ways of measuring meaning. A key concept which emerged is *semantic relatedness*, which seeks to quantify human intuitions such as: *tree* and *flower* are closer in meaning than *tree* and *hamburger*. Simple early approaches analysed term co-occurrence in large corpora [7, 17]. Later, more sophisticated approaches such as Latent Semantic Analysis constructed vectors around the compared terms (consisting of, for instance, word counts in paragraphs, or documents) and computed their cosine similarity.

Innovative extensions to these methods appeared following the recent explosion in free user-supplied Web content, including the astoundingly detailed and organized Wikipedia. Thus [6] enrich their term vectors with Wikipedia article text: an approach called Explicit Semantic Analysis. [14] develop a similar approach using only Wikipedia's internal hyperlinks. Here semantic relatedness effectively becomes a measure of likelihood that each term will be anchor text in a link to a Wikipedia article about the other.

In the background of this research lurk fascinating philosophical questions. Is closeness in meaning sensibly measured in a single numeric value? If not, how should it be measured? Can the semantic relatedness of two terms be measured overall, or does it depend on the context where they occur? Yet automated measures of semantic relatedness now have a high correlation with native human judgments [13].

4 AUTOMATED ONTOLOGY BUILDING: STATE OF THE ART

Dissatisfaction with the limitations of manual ontology-building projects such as Cyc led to a lull in formal knowledge representation through the 1990s and early 2000s, but the new methods of determining semantic relatedness described above, and the free user-supplied Web content on which they draw, has recently begun a new era in *automated* ontology building.

One of the earliest projects was YAGO [20, 21], which maps Wikipedia's leaf categories onto the WordNet taxonomy of synsets, adding articles belonging to those categories as new elements, then extracting further relations to augment the taxonomy. Much useful

information is obtained by parsing category names, for example extracting relations such as *bornInYear* from categories such as *1879 birth*.

A much larger, but less formally structured, project is DBpedia [1, 2], which transforms Wikipedia's infoboxes and related features into a vast set of RDF triples (103M), to provide a giant open dataset on the web. This has since become the hub of a Linked Data Movement which boasts billions of triples [3]. Due to the lack of formal structure there is however much polysemy and many semantic relationships are obscured (e.g. there are redundant relations from different infobox templates, for instance *birth_date*, *birth* and *born*). Therefore they have also released a DBpedia Ontology generated by manually reducing the most common Wikipedia infobox templates to 170 ontology classes and the 2350 template relations to 940 ontology relations asserted onto 882,000 separate instances.

The European Media Lab Research Institute (EMLR) built an ontology from Wikipedia's category network in stages. First they identified and isolated *isA* relations from other links between categories [16]. Then they divided *isA* relations into *isSubclassOf* and *isInstanceOf* [24], followed by a series of more specific relations (e.g. *partOf*, *bornIn*) by parsing category titles and adding facts derived from articles in those categories [15]. The final result consists of 9M facts indexed on 2M terms in 105K categories.[7]

What is notable about these projects is that firstly, all have found it necessary to build on a manually created backbone (in the case of YAGO: Wordnet, in the case of the EMLR project: Wikipedia's category network, and even DBPedia produced its own taxonomy). Yet none of these ontologies can recognize the wrongness of `S1`. Although YAGO and EMLR's system possess rich taxonomic structure, it is property-based rather than categorical, and does not enforce the relevant constraints. A second important issue concerns evaluation. With automation, accuracy becomes a key issue. Both YAGO and DBPedia (and Linked Data) lack any formal evaluation, though EMLR did evaluate the first two stages of their project — interestingly, using Cyc as a gold standard — reporting precision of 86.6% and 82.4% respectively.

Therefore we wondered whether Cyc's more stringent categorical knowledge might serve as an even more effective backbone for automated ontology-building, and also whether we might improve on the accuracy measurement from EMLR. We tested these hypotheses in a practical project, which transferred knowledge automatically from Wikipedia to Cyc (ResearchCyc version 1.0).

5 AUTOMATED ONTOLOGY BUILDING: CYC AND WIKIPEDIA

5.1 Stage 1: Concept mapping

Mappings were found using four stages:

Stage A: Searches for a one-to-one match between Cyc term and Wikipedia article title.

Stage B: Uses Cyc term synonyms with Wikipedia redirects to determine a single mapping.

Stage C: When multiple articles map, a 'context' set of articles (comprised of article mappings for Cyc terms linked to the current term) is used to identify the article with the highest semantic-related score using [14].

Stage D: Disambiguates and removes incorrect mappings by performing Stage A and B backwards (e.g. `DirectorOfOrganisation` → *Film director* → `Director-Film`, so this mapping is discarded).

[7] Downloadable at `http://www.eml-research.de/english/research/nlp/download/wikirelations.php`

5.2 Stage 2: Transferring knowledge

Here new subclasses and instances ('children') were added to the Cyc taxonomy, as follows.

5.2.1 *Finding possible children*

Potential children were identified as articles within categories where the category had an equivalent Wikipedia article mapped to a Cyc collection (about 20% of mapped articles have equivalent categories).

Wikipedia's category structure is not as well-defined as Cyc's collection hierarchy, containing many merely associatively-related articles. For example *Dogs* includes *Fear of dogs* and *Puppy Bowl*. Blind harvesting of articles from categories as subclasses and instances of Cyc concepts was therefore inappropriate.

5.2.2 *Identifying correct candidate children*

Each article within the given category was checked to see if a mapping to it already existed from a Cyc term. If so, the Cyc term was taken as the child, and the relevant assertion of parenthood made if it did not already exist. If not, a new child term was created if verified by the following methods:

Link parsing: The first sentence of an article can identify parent candidates by parsing links from a regularly structured sentence. Each link represents a potential parent if the linked articles are already mapped to Cyc collections (in fact multiple parents were identified with this method).

The regular expression set was created from the most frequently occurring sentence structures seen in Wikipedia article first sentences. Examples included:

- *X are a Y*
  'Bloc Party are a British indie rock band…'
- *X is one of the Y*
  'Dubai is one of the seven emirates…'
- *X is a Z of Y*
  'The Basque Shepherd Dog is a breed of dog…'
- *X are the Y*
  'The Japanese people are the predominant ethnic group of Japan.'

Infobox pairing: If an article within a category was not found to be a child through link parsing, it was still asserted as a child if it shared the same infobox template as 90% of the children that *were* found.

5.2.3 *Results*

The project added over 35K new concepts to the lower reaches of the Cyc ontology, each with an average of seven assertions, effectively growing it by 30%. It also added documentation assertions from the first sentence of the relevant Wikipedia article to the 50% of mapped Cyc concepts which lacked this, as illustrated in Figure 2.

An evaluation of these results was performed with 22 human subjects on testsets of 100 concepts each. It showed that the final mappings had 93% precision, and that the assignment of newly created concepts to their 'parent' concepts was 'correct or close' 90% of the

Figure 2. A Cyc concept containing information added from Wikipedia.

time [18]. This suggests a modest improvement on the EMLR results, though more extensive testing would be required to prove this. Work on an earlier version of the algorithm [12] also tested its accuracy against the inter-agreement of six human raters, measuring the latter at 39.8% and the agreement between algorithm and humans as 39.2%.

5.3 Categorical quality control

During the initial mapping stage, Cyc's disjointness knowledge was put to work discriminating rival candidate matches to Cyc concepts which had near-equal scores in quantitative semantic relatedness. In such cases Cyc was queried for disjointness between ancestor categories of the rivals, and if disjointness existed, the match with the highest score was retained and others discarded. Failing that, all high-scoring matches were kept. Examples of where this worked well were the Wikipedia article *Valentine's Day*, which mapped to both `ValentinesDay` and `ValentinesCard`, but Cyc knew that a card is a spatiotemporal object and a day is a 'situation', so only the former was kept. On the other hand, the test allowed *Black Pepper* to be mapped to both `BlackPeppercorn` and `Pepper-TheSpice`, which despite appearances was correct given the content of the Wikipedia article.

During the knowledge transfer stage an interesting phenomenon occurred. Cyc was insistently 'spitting out' a given assertion and it was thought that a bug had occurred. To the researchers' surprise it was found that Cyc was ontologically correct. From that time on, the assertions Cyc was rejecting were gathered in a file for inspection. At the close of the project this file contained 4300 assertions, roughly 3% of the assertions fed to Cyc. Manual inspection suggested that 96% of these were 'true negatives,' for example:

```
(isa CallumRoberts Research)
(isa Insight-EMailClient EMailMessage)
```

This compares favourably with the evaluated precision of assertions successfully added to Cyc.

The examples above usefully highlight a clear difference between quantitative measures of semantic relatedness, and an ontological relatedness derivable from a principled category structure. Callum Roberts is a *researcher*, which is highly semantically related to *research* and Insight is an *email client*, which is highly semantically related to *email messages*. Thematically or topically these pairs are incredibly close, but ontologically speaking, they are very different kinds of thing. Thus if we state:

```
S4: Callum Roberts is a research
```

we once again hit the distinctively unsettling silliness of the traditional philosophical category mistake, and a kind of communication we wish our computers to avoid.

6 PLANS FOR FURTHER FEEDING

Given the distinction between semantic and ontological relatedness, we may note that combining the two has powerful possibilities. In fact this observation may usefully be generalized to note that in automated information science, *overlapping independent heuristics* are a boon to accuracy, and this general principle will guide our research over the next few years.

Our first step will be to develop strategies to automatically augment Cyc's disjointness network and semantic argument constraints on relations (where Cyc's manual coding has resulted in excellent precision but many gaps) using features from Wikipedia. For instance, systematically organized infobox relations, helpfully collected in DBPedia, are a natural ground to generalize argument constraints. The Wikipedia category network will be mined — with caution — for further disjointness knowledge. This further commonsense categorical knowledge will then bootstrap further automated ontology-building.

7 PHILOSOPHICAL LESSONS

Beyond the practical results described above, our project provides fuel for philosophical reflection. It suggests the notion of philosophical categories should be rehabilitated as it leads to measurable improvements in real-world ontology-building. Just how extensive a system of categories should be will of course require real-world testing. But now we have the tools, the computing power, and most importantly the wealth of free user-supplied data to do this. The issue of where exactly the line should be drawn between categories proper and mere properties remains open. However, modern statistical tools raise the possibility of a quantitative treatment of ontological relatedness that is more nuanced than Aristotle's ten neat piles of predicates, yet can still recognize that `S1` is highly problematic, and why.

REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, 'Dbpedia: A nucleus for a web of open data', *The Semantic Web*, 722–735, (2007).

[2] S. Auer and J. Lehmann, 'What have innsbruck and leipzig in common? extracting semantics from wiki content', *The Semantic Web: Research and Applications*, 503–517, (2007).

[3] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, 'Dbpedia-a crystallization point for the web of data', *Web Semantics: Science, Services and Agents on the World Wide Web*, **7**(3), 154–165, (2009).

[4] R. Carnap, 'The elimination of metaphysics through logical analysis oj language', *Erkenntnis*, **2**(1), 219–241, (1931).

[5] R. M. Chisholm, *A realistic theory of categories: An essay on ontology*, volume 146, Cambridge Univ Press, 1996.

[6] E. Gabrilovich and S. Markovitch, 'Computing semantic relatedness using wikipedia-based explicit semantic analysis', in *Proceedings of the 20th international joint conference on artificial intelligence*, volume 6, p. 12. Morgan Kaufmann Publishers Inc., (2007).

[7] J. J. Jiang and D. W. Conrath, 'Semantic similarity based on corpus statistics and lexical taxonomy', in *Proc. of the Int'l. Conf. on Research in Computational Linguistics*, pp. 19–33, (1997).

[8] D. B. Lenat, 'Cyc: A large-scale investment in knowledge infrastructure', *Communications of the ACM*, **38**(11), 33–38, (1995).

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

[10] E. J. Lowe, 'Ontological categories and natural kinds', *Philosophical papers*, **26**(1), 29–46, (1997).

[11] E. J. Lowe, *The possibility of metaphysics: Substance, identity, and time*, Oxford University Press, USA, 2001.

[12] O. Medelyan and C. Legg, 'Integrating cyc and wikipedia: Folksonomy meets rigorously defined common-sense', in *Proceedings of the WIKI-AI: Wikipedia and AI Workshop at the AAAI*, volume 8, (2008).

[13] O. Medelyan, D. Milne, C. Legg, and I. H. Witten, 'Mining meaning from wikipedia', *International Journal of Human-Computer Studies*, **67**(9), 716–754, (2009).

[14] D. Milne and I. H. Witten, 'An effective, low-cost measure of semantic relatedness obtained from wikipedia links', in *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, (2008).

[15] V. Nastase and M. Strube, 'Decoding wikipedia categories for knowledge acquisition', in *Proceedings of the 23rd national conference on Artificial intelligence*, volume 2, pp. 1219–1224, (2008).

[16] S. P. Ponzetto and M. Strube, 'Deriving a large scale taxonomy from wikipedia', in *Proceedings of the national conference on artificial intelligence*, volume 22, p. 1440. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, (2007).

[17] P. Resnik, 'Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language', *Journal of Artificial Intelligence Research*, **11**, 95–130, (1999).

[18] S. Sarjant, C. Legg, M. Robinson, and O. Medelyan, "all you can eat' ontology-building: Feeding wikipedia to cyc', in *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pp. 341–348. IEEE Computer Society, (2009).

[19] M. Schlick, 'Meaning and verification', *The philosophical review*, **45**(4), 339–369, (1936).

[20] F. M. Suchanek, G. Kasneci, and G. Weikum, 'YAGO: a core of semantic knowledge', in *Proceedings of the 16th international conference on World Wide Web*, pp. 697–706. ACM, (2007).

[21] F. M. Suchanek, G. Kasneci, and G. Weikum, 'YAGO: A large ontology from wikipedia and wordnet', *Web Semantics: Science, Services and Agents on the World Wide Web*, **6**(3), 203–217, (2008).

[22] W. van Orman Quine, 'On what there is', *From a Logical Point of View*, 1–19, (1953).

[23] P. Weiss, *Modes of being*, volume 2, Southern Illinois Univ Pr, 1958.

[24] C. Zirn, V. Nastase, and M. Strube, 'Distinguishing between instances and classes in the wikipedia taxonomy', in *Proceedings of the 5th European semantic web conference on The semantic web: research and applications*, pp. 376–387. Springer-Verlag, (2008).

Synthetic Semiotics: on modelling and simulating the emergence of sign processes

Angelo Loula[1] **& João Queiroz**[2]

Abstract. Based on formal-theoretical principles about the sign processes involved, we have built synthetic experiments to investigate the emergence of communication based on symbols and indexes in a distributed system of sign users, following theoretical constraints from C.S.Peirce theory of signs, following a Synthetic Semiotics approach. In this paper, we summarize these computational experiments and results regarding associative learning processes of symbolic sign modality and cognitive conditions in an evolutionary process for the emergence of either symbol-based or index-based communication.

1 INTRODUCTION

Following the motto 'build to explain', a synthetic approach (opposed to an analytical one) corresponds to a reverse methodology that builds creatures and environments describing a simple and controllable framework to generate, test and evaluate theories and hypothesis about the system being modelled. Diverse processes and systems are modelled and simulated in such synthetic experiments, including social, biological and cognitive processes [1, 2, 3, 4, 5, 6]. Particularly, we have been modelling and simulating semiotic systems and processes, following a Synthetic Semiotics approach.

Based on formal-theoretical principles about the sign processes involved, we have built synthetic experiments to investigate the emergence of communication based on symbols and indexes in a distributed system of sign users, following theoretical constraints from C.S.Peirce theory of signs. In this paper, we summarize these computational experiments and results. We investigated the associative learning processes of symbolic sign modality and the relation between different sign modalities in the transition from indexical to symbolic communication. We also studied cognitive conditions in an evolutionary process for the emergence of either symbol-based or index-based communication, relying on different types of cognitive architecture.

First, we review related work, then we describe our formal-theoretical background, the sign theory by of C.S.Peirce. Finally we present synthetic experiments that modelled and simulated the emergence of communication processes, dealing with the learning process of symbolic sign modality and also with the evolution of indexical and symbolic interpretative behaviours. The notion of responsive environments is broad, encompassing essentially every space capable of sensing and responding accordingly to entities that inhabit them (these entities can be people, animals, or any sort of identifiable objects).

2 RELATED WORK

There have been several different experiments concerning symbol grounding and also the emergence of shared vocabularies and language in simple (real or virtual) worlds [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18] (for a review of other works, see [19], [20]). Despite the fact that sign processes are in the foundations of communication, little discussion about such processes can be found, such as the emergence of fundamental types of signs and their interpretative effects.

There have been studies introducing theoretical foundations in reference to Peirce's work [11, 16, 17, 13, 8], but they just borrow Peircean definitions of symbol or of sign without generating any further consequences to the designed experiment. For example, in [17], [13] and [8], authors bring forth definitions of signs and symbols from Peirce's general theory of signs, but they end up changing them, in such a way that it is not possible to conclude whether the experiments were actually based on Peirce's theory or whether it contributed, validating it or not, some of the principles of Peirce's theory. In [11] and [16], on the other hand, presents Peirce's theory through a second hand reading of Deacon's work, which is at least a limited analysis of the Peircean theory and, in special, of his definition of a symbol. As a consequence, we can say that they were not able to recognize a symbol when it first occurred in their experiments.

Deacon's reading of Peirce's theory is the most popular example at hand of such disconnection between theoretical framework and actual research [21]. His depiction of humans as the only 'symbolic species' is based on the assumption that symbols necessarily have combinatory properties, and that only the human prefrontal cortex could possibly implement such properties. However, this proposal is incongruent with Peirce's theory and frontally collides with several empirical lines of evidence (for a discussion of this point, see [22],[23]). Poeppel [24] already recognized the 'problematic' and 'speculative' manner in which Deacon built his arguments using Peirce's theory, comparative and evolutionary approaches to language and even linguistic theories.

We claim that just bringing forward a definition from Peirce's theory without deriving any consequence or constraint to the experimental setup certainly reduces the explanatory power of the proposed model. Recognizing the inter-dependence of Peirce's concepts at different levels, such as the sign model and its derived sign classification, substantially enriches computational experiments willing to simulate communication and its relationship to meaning.

[1] Intelligent and Cognitive Systems Lab, State University of Feira de Santana, Brazil. Email: angeloc1@ecomp.uefs.br.

[2] Institute of Arts and Design; Federal University of Juiz de Fora, Brazil. Email: queirozj@pq.cnpq.br.

3 THE THEORY OF SIGNS OF C.S. PEIRCE

North-American pragmatist Charles Sanders Peirce, founder of the modern theory of signs, defined semiotics as a kind of logic: a science of the essential and fundamental nature of all possible varieties of meaning processes (*semiosis*). Peirce's concept of semiotics as the 'formal science of signs', and the pragmatic notion of meaning as the 'action of signs', have had a deep impact in philosophy, in theoretical biology and in cognitive science (see [25]). Peircean approach to semiotic process (semiosis) is also related to formal attempts to describe cognitive processes in general. His framework provides: (i) a list of fundamental varieties of representations based on a theory of logical categories; (ii) a model to approach the emergence and evolution of semiotic complexity in artificial and biological systems.

Peirce defined semiosis (meaning process) as an irreducible triadic relation between a sign (S), its object (O) and its interpretant (I). That is, according to Peirce, any description of semiosis involves a relation constituted by three irreducibly connected terms: "A sign is anything which determines something else (its interpretant) to refer to an object to which [it] itself refers (its object) in the same way, the interpretant becoming in turn a sign, and so on ad infinitum" [26, CP 2.303]. Semiosis is also characterized as a behavioural pattern that emerges through the intra/inter-cooperation between agents in a communication act, which involves an utterer, a sign, and an interpreter. Meaning and communication processes are defined in terms of the same "basic theoretical relationships" [27], i.e., in terms of a self-corrective process whose structure exhibits an irreducible relation between three elements. In a communication process, "[i]t is convenient to speak as if the sign originated with an utterer and determined its interpretant in the mind of an interpreter" [28, MS 318].

As it is well known, sign-mediated processes show a notable variety. There are three fundamental kinds of signs underlying meaning processes – icons, indexes, and symbols [26, CP 2.275]. They correspond to similar, reactive, and law relationship which can be established between a sign and its object. Icons are signs that stand to objects by similarity, without regard to any space-time connection with existing objects [26, CP 2.299]. An icon stands to the object independently of any spatio-temporal presence of the latter; it refers to the object merely by virtue of its own properties. This is an important feature distinguishing iconic from indexical sign-mediated processes. Indices are signs that refer to objects due to a direct physical connection between them. Accordingly, spatio-temporal co-variation is the most characteristic aspect of indexical processes. Finally, symbols are signs that are related to their object through a determinative relation of law, rule or convention. A symbol becomes a sign of some object merely or mainly by the fact that it is used and understood as such.

4 EXPERIMENTS IN SYNTHETIC SEMIOTICS

4.1 Learning and the emergence of symbol-based communication

Inspired by the vervet monkey alarm call ethological study case ([29], see [23], for a neurosemiotic analysis), we have simulated an ecosystem for artificial creatures' interactions, including intra-specific communication for predators' presence. We investigated the learning processes (habit acquisition) of symbolic sign modality and the relation between different sign modalities in the transition from indexical to symbolic behaviour through associative learning.

The creatures were autonomous agents inhabiting a virtual bi-dimensional environment. This virtual world was composed of prey and predators (terrestrial, aerial and ground predators), and of things such as trees (climbable objects) and bushes (used to hide). Preys could produce vocalizations (alarm calls) indicating that a predator was seen. That vocalization could become immediately available to nearby preys by way of a hearing sensor. We proposed two scenarios: with apprentices and tutors [30], and with self-organizers [31]. Apprentices and tutors, as seen in the contrast between infant and adult vervet monkeys, defined a learning relation. Tutors, that had already established vocalizations for each predator, were the only ones to vocalize and as the preys heard them, they tried to establish the connections relations between the auditory and the visual stimuli. Self-organizer creatures were apprentices and tutors at the same time, but there was no initially established repertoire of alarms calls, and the group of preys had to create and share alarm calls for each predator, by vocalizing to and learning from each other.

Associative learning was the mechanism used by preys to gradually acquire association rules between auditory and visual data necessary to interpret signs as symbols. It involved working memories and an associative memory. Working memories allows the persistence of spatio-temporal relations. Associative memory formation followed Hebbian learning principles [32] and allowed the creatures to, not only, learn temporal and spatial relations from the external stimuli and the associations to be created, but also reinforced or weakened them (varying association strength between 0 and 1) according to the co-occurrence of stimuli in the working memories (figure 1).

After hearing a vocalization, preys initially responded with a sensorial scan for the utterer and co-occurring events, a typical indexical behaviour. As the strength of sign-predator associations reached a certain threshold, after multiples reinforcements, a new action rule was established, 'flee with no scanning'. In this situation, the prey used an established association to interpret the alarm, and we can say that the sign-object relation depended on the interpreter and no more in a physical, spatial-temporal evidence, and therefore the alarm became a symbol.

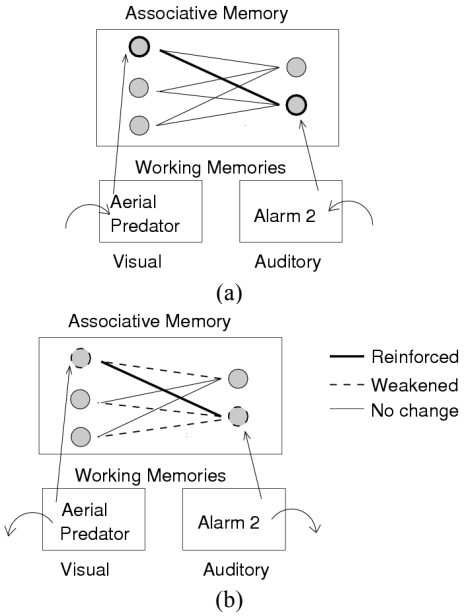


Figure 1. Associative learning: reinforcement and weakening. (a) The co-occurrence of visual and auditory stimuli in working memories reinforces the association between them. (b) When sensory stimuli are dropped from working memories, associations involving them and that were are not reinforced, are weakened.

During simulations, we observed the associative memory items and behaviour responses of the preys to alarm calls. Results showed that both apprentice and self-organizer preys were able to acquire symbolic competence. Preys initially exhibited an indexical behaviour to alarm calls, but a symbolic response emerged by means of communicative interactions. Apprentices were able to establish the same alarm-predator relations used by tutors (alarm 1 - terrestrial predator, alarm 2 - aerial predator, alarm 3 - ground predator). Even though apprentices, eventually associated alarms with the presence of elements such as trees and bushes, the associative learning mechanism was able to gradually reinforce the correct links, going up to its maximum value of 1.0 at the end of simulation, while weakening the other links, which went down the minimum value of zero (figure 2; see [30], for more detailed results).

On the other side, self-organizers, starting with no a priori relation between alarms and predators, were able, at the end, to converge to a common repertoire of associations between alarms and predators. As there were no predefined alarms for each predator, each creature could create a random alarm (from 0 to 99) for a predator if it had not had one associated with that predator before. As a consequence, various alarms were created for the same predator, and even the same alarm could be used for different predators. And some alarms could also be associated with elements other than predators. Nevertheless, associative learning was responsible for a gradual convergence of the community of preys to use the same alarms for the same predators (figure 3; see [31], for more detailed results).

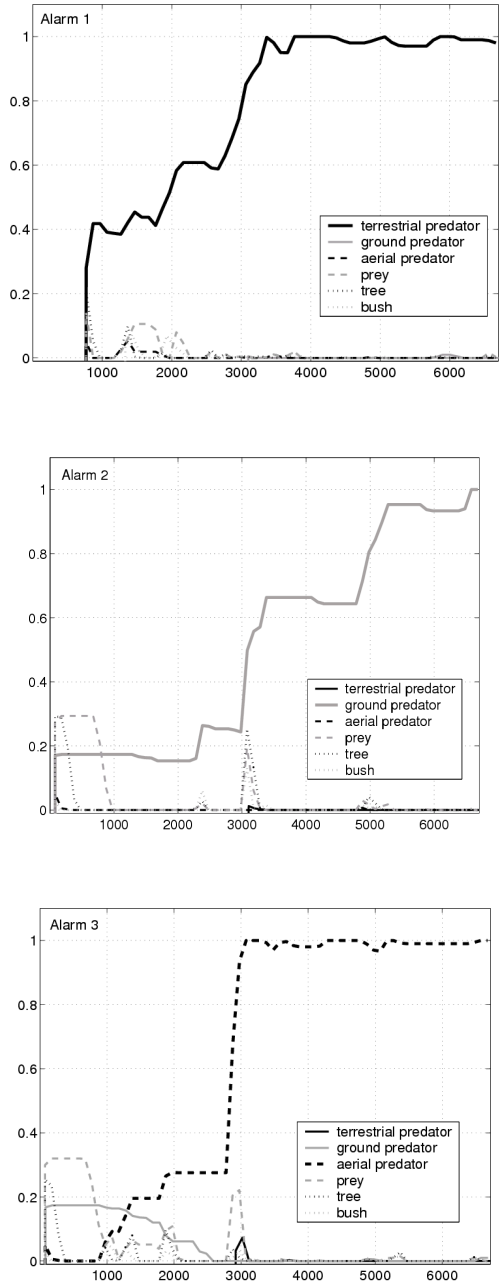


Figure 2. Associations' strength values for one apprentice, for each alarm, during simulation.

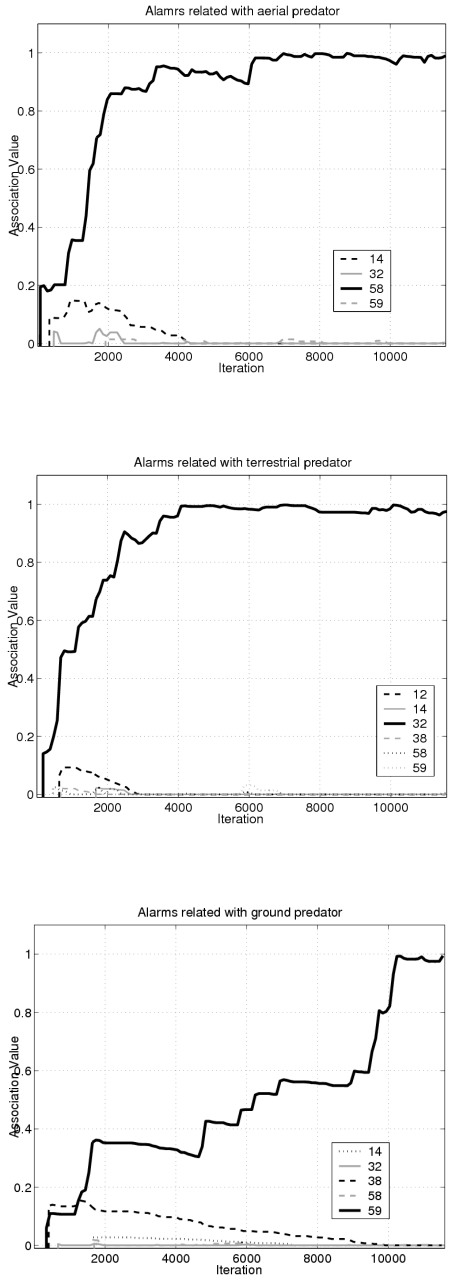


Figure 3. Associations' strength values for self-organizers, for each type of predator.

4.2 Evolution and the emergence of different classes of sign processes

Based on the fact that signs can be of different types and that communication processes rely on the production and interpretation of signs, we have modeled the emergence of indexical and symbolic interpretative behaviors in communication processes, when none of them was initially available, and we have studied how they emerge, and the cognitive conditions for the emergence of such interpretation processes. To model those interpretation-communication processes, we also followed the minimum brain model for vocalization behavior in from [23] and the biological motivations from animal communication, specifically, for food calls [33].

Indexical interpretation is a reactive interpretation of signs, so for our creatures to have this competence, they had to be able to reactively respond to sensory stimulus with prompt motor answer. But then again a symbolic interpretation undergoes the mediation of the interpreter to connect the sign to its object, in such a way that a habit (either inborn or acquired) must be present to establish this association. Also, in symbolic interpretation, an associative memory must be present as it is the only domain able to establish connections between different representation modes. Thus, our artificial creatures had to be able to receive sensory data, both visual and auditory, that could be connected directly to motor responses (Type 1 architecture), or else they should be connected to motor responses indirectly, through the mediation of an associative memory, that associates auditory stimulus to visual stimulus (Type 2 architecture) (see figure 4).

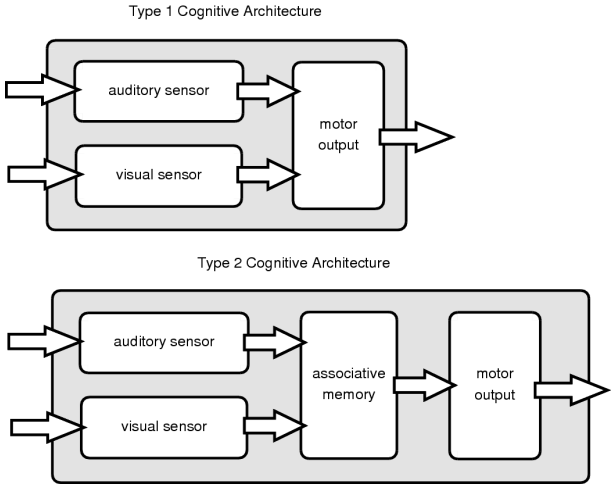


Figure 4. Cognitive architectures for representations' interpretations. Top: Type 1 architecture. Bottom: Type 2 architecture.

Lower quality resources were scattered throughout the environment and a single location received highest quality resources, where one creature (vocalizer) was placed. The other creatures (interpreters) were controlled by finite state machines

(FSM) and had visual and auditory sensors and motor capabilities. These interpreter creatures could respond to visual inputs with one of the motor actions, and could also respond to auditory input with a direct motor action (a reactive, indexical process) (Type 1 architecture). Alternatively, before an input was sent to the FSM, they could also choose to establish an internal association between the heard stimulus and the visual representation domain (Type 2 architecture). This internal association linked what was heard with the view of a collectible resource, i.e. the creature could interpret the sign heard as a resource and act as if the resource was seen.

At the start of the simulations, interpreter creatures were randomly defined, so creatures did not respond appropriately to sensory inputs. But an evolutionary process of variation and selection was applied, allowing the evolution of individuals to better accomplish the task of resource foraging. During the evolutionary process, for each start-up conditions, we observed the types of cognitive architecture used by creatures and their motor responses to sensory input.

We performed two initial experiments to evaluate the emergence of either an indexical interpretation or a symbolic interpretation of vocalizations. Such experiments involved 2 cycles, but only in the second cycle, the vocalizer was present. In the first experiment, creatures just had to have a specified action as output of the FSM to execute that action. We observed that the indexical interpretation was the competence acquired by creatures to deal with communication, with the direct association between auditory signs and motor actions. But, in a second experiment, for motor actions to be executed, the creatures needed to first output a null action before any movement action was done. In this case, learning motor coordination was harder. In this alternative scenario, symbolic interpretation was the emerging competence, instead of an indexical one. We asserted the hypothesis that acquiring symbolic competence would act as a cognitive shortcut, by reusing a previously acquired ability in cycle 1 to appropriately respond to visual data with motor actions. We proposed that a symbolic interpretation process can happen if a cognitive trait is hard to be acquired and the symbolic interpretation of a sign will connect it with another sign for which the creature already has an appropriate response (figure 5; see [34] for detailed results).

Once symbolic interpretation needed a competence to benefit from, we investigated the availability and reliability of such previous competence in a subsequent set of experiments. We first proposed an experiment where this first cycle did not occur, therefore visual-motor coordination was not established before vocalizations started. From this single cycle experiment, it was possible to observe that even though the vocalizer was available from start, creatures did not use signs at all in a first moment. But, as trying to acquire visual-motor coordination and also a sign-motor coordination was a hard task route, the symbolic interpretation diminished this effort and became the dominant strategy (figure 6; see [35], for more detailed results).

To go further in our investigation, we set up another experiment, in which cycle 1 was present but there was a failure chance in the visual-motor coordination after cycle 1, simulating a malfunctioning cognitive module. At first, with a 20% of motor action selection failure, symbolic processes were still established, with reuse of a degraded module, with a relative increase in foraging efficiency, however. A higher failure of 50% proved to worsen the performance of the visual control module considerably more, and allowed indexical interpretation of sign to be established, as a way to avoid reusing it.

At the end of our experiments, we confirmed our hypothesis that symbolic competence acted as a cognitive shortcut, and, as such, the cognitive module to which the symbolic interpretation was connecting to must be already established. Nevertheless, it does need to be fully functional, as minimal visual-motor coordination is sufficient to begin a symbolic interpretation process and even a moderately damaged module can also be reused.

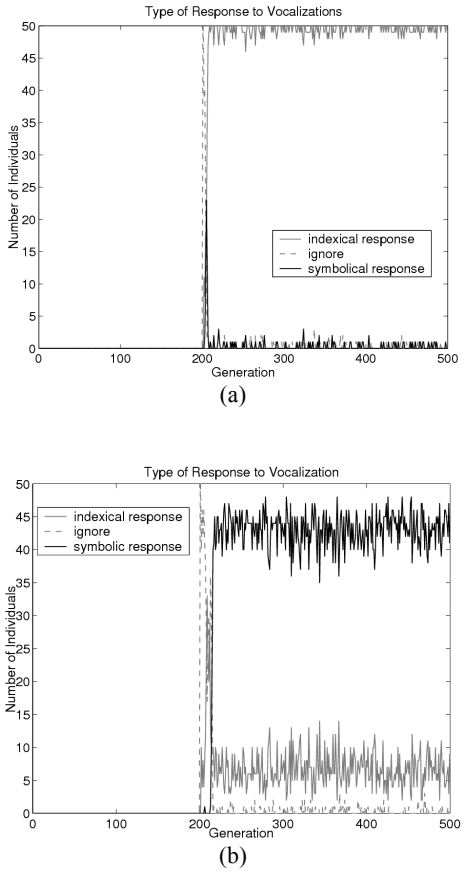
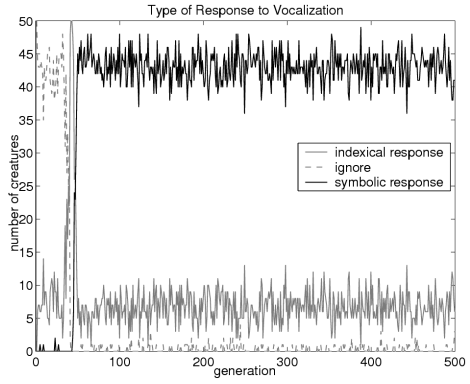
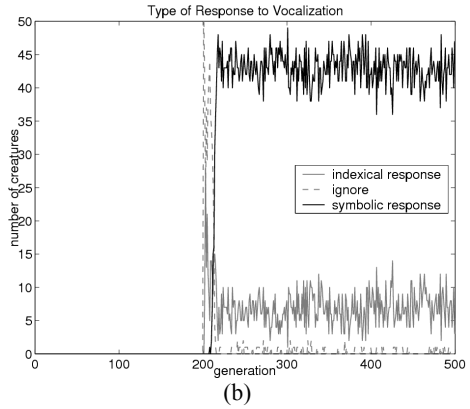


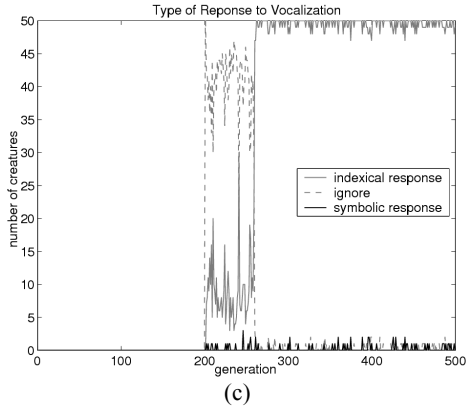
Figure 5. Evaluation of the type of response to vocalizations along the generations for (a) the direct motor action experiment and (b) the previous null action experiment.



(a)



(b)



(c)

Figure 6. Evaluation of the type of response to vocalizations along the generations for (a) the one cycle only experiment, (b) 20% failure experiment, and (c) 50% failure experiment.

5 CONCLUSIONS

The relation between simulations and theories is a 'two-way road' (see [36]). Simulations offer to the theory an opportunity to formalize and quantify, in terms of programming language. Following a synthetic approach, a computational model is built based on basic theoretical assumptions about the target-system. Here, we applied the sign theory of C.S.Peirce in building synthetic experiments to investigate the transition to symbolic communication by means of associative learning and cognitive conditions in a evolutionary processes for either symbol-based communication or index-based communication.

Even though Peirce's pragmatic approach have established a rigorous distinction between different classes of sign processes as well as between semiotic behaviour and brute reactive behaviour, he did not describe: (i) the dynamics responsible for the emergence of semiosis in an evolutionary scenario, and (ii) the dynamics responsible for the transition from iconic and indexical semiotic systems to symbolic and meta-semiotic ones. Synthetic Semiotics can define a methodology for better understanding the dynamics related to the emergence of indexical and symbolic-based semiosis. Formal-theoretical principles act not only as theoretical background but also as constraints in designing the artificial systems and as bridges for contributions to the sign theory that originally provided the principles.

REFERENCES

[1] V. Braitenberg, *Vehicles - Experiments in Synthetic Psychology*. Cambridge, MA: MIT Press, 1984.

[2] C. Langton, editor. *Artificial Life: an overview*. MIT Press, 1995.

[3] J. Noble, 'The scientific status of artificial life', In Fourth *European Conference on Artificial Life (ECAL97)*, Brighton, UK, (1997).

[4] T. Froese and T. Ziemke, 'Enactive artificial intelligence: Investigating the systemic organization of life and mind', *Artificial Intelligence*, 173, 466–500, (2009).

[5] R. Pfeifer, F. Iida, and J. Bongard, 'New robotics: Design principles for intelligent systems', *Artificial Life*, 11 (1-2), 99–120, (2005).

[6] R. Brooks, 'Intelligence without reason', In *Proceedings of the 12th International Joint Conference on Artificial Intelligence - IJCAI-91*, pp. 569–595, San Mateo, CA: Morgan Kauffmann, (1991).

[7] D. Roy, 'Grounding Words in Perception and Action: Insights from Computational Models', *Trends in Cognitive Science*, 9 (8): 389-96, (2005).

[8] D. Roy, 'Semiotic Schemas: A Framework for Grounding Language in the Action and Perception', *Artificial Intelligence*, 167 (1-2): 170-205, (2005).

[9] L. Steels, *The Talking Heads Experiment: Volume I. Words and Meanings*. VUB Artificial Intelligence Laboratory, Brussels, Belgium. Special pre-edition, (1999).

[10] L. Steels 'Evolving grounded communication for robots'. *Trends Cogn. Sci.* 7, 308-312, (2003).

[11] A. Cangelosi, A. Greco, and S. Harnad. 'Symbol grounding and the symbolic theft hypothesis'. In A. Cangelosi and D. Parisi, editors, *Simulating the Evolution of Language* (chap.9). London:Sprinter, (2002).

[12] A. Cangelosi and H. Turner. 'L'emergere del linguaggio'. In A. M. Borghi and T. Iachini, editors, *Scienze della Mente,* pp.227-244, Bologna: Il Mulino, (2002).

[13] P. Vogt. 'The physical symbol grounding problem'. *Cognitive Systems Research*, 3(3), 429–457, (2002).
[14] B. J. MacLennan, 'Synthetic ethology: a new tool for investigating animal cognition'. In *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition*, ch.20, pp.151-156, Cambridge, Mass.: MIT Press, (2002).
[15] B. J. MacLennan. 'The emergence of communication through synthetic evolution'. In *Advances in the Evolutionary Synthesis of Intelligent Agents*, pp. 65-90, Cambridge, Mass.: MIT Press, (2001).
[16] D. Jung and A. Zelinsky. 'Grounded symbolic communication between heterogeneous cooperating robots'. Autonomous Robots journal, 8(3), 269–292, (2000).
[17] R. Sun, 'Symbol grounding: A new look at an old idea'. *Philosofical Psychology*, 13(2), 149–172, (2000).
[18] E. Hutchins and B. Hazlehurst. 'How to invent a lexicon: the development of shared symbols in interaction'. In *Artificial Societies: The Computer Simulation of Social Life*. London: UCL Press, (1995).
[19] M.H. Christiansen and S. Kirby. Language evolution: consensus and controversies. *Trends in Cognitive Sciences*, 7 (7), 300-307, (2003).
[20] K. Wagner, J. A. Reggia, J. Uriagereka, and G. S. Wilkinson, 'Progress in the simulation of emergent communication and language'. *Adaptive Behavior*, 11(1):37—69, (2003).
[21] T. Deacon. *Symbolic Species: The Co-evolution of Language and the Brain*. New York: Norton, 1997.
[22] S. Ribeiro, A. Loula, I. Araújo, R. Gudwin, R. and J. Queiroz Symbols are not uniquely human. *Biosystems* 90(1): 263-272, (2007).
[23] J. Queiroz and S. Ribeiro 'The biological substrate of icons, indexes, and symbols in animal communication: A neurosemiotic analysis of vervet monkey alarm calls'. In *The Peirce Seminar Papers* 5, pp.69–78, Berghahn Books, New York, (2002).
[24] D. Poeppel. 'Mind over chatter'. *Nature* 388:734, (1997).
[25] J. Queiroz and F. Merrell. 'On Peirce´s pragmatic notion of semiosis – a contribution for the design of meaning machines'. *Minds & Machines* 19, 129-143, (2009).
[26] C.S. Peirce, *The collected papers of Charles Sanders Peirce*. Electronic edition. Vols.I-VI. C. Hartshorne and P. Weiss, editors. Charlottesville: Intelex Corporation. MA: Harvard University, 1931-1935. (cited using CP followed by volume number and page number)
[27] J. Ransdell. 'Some leading ideas of Peirce's semiotic'. *Semiotica*, 19, 157–178, (1977).
[28] C.S. Peirce. *Annotated catalogue of the papers of Charles S. Peirce*. R. Robin, editor. Amherst: University of Massachusetts, 1967. (cited using MS followed by manuscript number)
[29] D. L. Cheney and R. M. Seyfarth, *How monkeys see the world: Inside the mind of another species*. Chicago: University of Chicago Press, 1990.
[30] A. Loula, R. Gudwin, and J. Queiroz, 'Symbolic communication in artificial creatures: an experiment in artificial life'. *Lecture Notes in Computer Science*, 3171, 336–345, Advances in Artificial Intelligence - SBIA 200,. (2004).
[31] A. Loula, R. Gudwin, C. El-Hani, and J. Queiroz, 'Emergence of self-organized symbol-based communication in artificial creatures'. *Cognitive Systems Research*, 11(2), 131–147. (2010).
[32] D.O. Hebb *The Organization of Behavior: A Neuropsychological Theory*. John Wiley & Sons, New York, 1949.
[33] M. D. Hauser. *The Evolution of Communication*. Cambridge, MA: MIT Press, 1997.
[34] A. Loula, R. Gudwin, and J. Queiroz. 'On the emergence of indexical and symbolic interpretation in artificial creatures, or What is this I hear?' In Fellermann, H., et al., editors, *Artificial Life XII*, pages 862--868. MIT Press. (2010)
[35] A. Loula, R. Gudwin, and J. Queiroz. 'Cognitive conditions to the emergence of sign interpretation in artificial creatures'. In: Tom Lenaerts et al., *Proceedings of the 11th European Conference on the Synthesis and Simulation of Living Systems*, p. 497-504. MIT Press, (2011)
[36] D. Parisi, *Simulazioni - la realtà rifatta nel computer*. Bologna: Il Mulino, 2001.

EMERGENTISM, COMPUTER SIMULATIONS AND THE NOUMENAL

Emanuele Ratti[1]

Abstract. The theory of emergentism, in its main subdivision (i.e. epistemological and ontological emergentism), is arguably an useful example in order to explain the concept of antinomy of Kant's Pure Reason (a problem that is current also in contemporary philosophy), that is a thesis an antithesis that are both reasonable but irreconcilable. In using a thought experiment concerning the new technology of computer simulation, I shall show that one case in which we can solve the antinomy posed by epistemological/ontological emergentism is when we take into consideration the domain of A-Life.

1 ONTOLOGICAL AND EPISTEMOLOGICAL EMERGENTISM

There are plenty of definition of emergentism. Some involves properties [7] [9], some phenomena [1], some objects [8] [10]. This is not of interest here to understand which ones we should use, whether phenomena, objects or properties. What we assume here is that there is a common way to characterize the problem of emergentism. The main problem is the nature of the relation between the parts of a system and the system taken as a whole, i.e. a parts-whole relation problem . After having highlighted this, it is worth to point out that there is a common characterization of the different varieties of emergentism: it is said that there are an epistemological emergentism and an ontological emergentism.

In the ontological emergentism, the world is arranged in different levels. In the classical conception of ontological emergentism, (see Broad) at each level there are new entities, that are irreducible, inexplainable and unpredictable with respect to the lower level entities. One may argue that this is a kind of naïve levelism, and it is too much for ontological emergentism. Morgan [10] outlines a metaphysical model of emergentism whilst Kim [8], analyzes it from the point of view of contemporary philosophical problems. What emerges from this analysis is that the metaphysical model of emergentism is exactly a type of 'levelism', i.e. the world is characterized by different levels, hierarchically organized from the lowest to the highest, based on increasing complexity. This is not a metaphor; this a precise ontological top-down structure, so that these levels are not merely levels of description:

"The world as portrayed in the new picture consists of an array of levels, each level consisting of two components: a set of entities constituting the domain of particulars for that level and a set of properties defined over this domain", [7, p.190].

The 'array' is given by mereological relations, i.e. the relation of 'being part of' so that the entities at an higher level B are composed of the entities of the lower level A. But in ontological emergentism, there may be that entities of A are not sufficient in order to obtain the entities of B (it is not needed that this happens for each object of each level).

So, as I pointed out, ontological emergentism is a kind of levelism and at each level, there are entities that are composed by parts, in the sense that each of these entities is composed by entities of the lower level. It may happen that some of the entities of the higher level, even if they are composed by parts, they are more than the sum of these parts, in the sense that they are irreducible and they are ontologically novel (not merely novel at a level of description) with respect to the entities of the lower level. The entities of this lower level are wholes too, they are composed by entities of the subsequent lower level, but they may be novel too in respect with the entities of the subsequent lower level. We can go downward until we find the level of physics, in which the entities have no proper parts. In Silbertstein and McGeever's words:

"Ontologically emergent features are neither reducible to nor determined by more basic features. Ontologically emergent features are features of systems or wholes that possess causal capacities not reducible to any of the intrinsic causal capacities of the parts nor to any of the (reducible) relations between the parts. Ontological emergence entails the failure of part–whole reductionism in both its explicit and mereological supervenience forms. It should be noted that epiphenomenal features do not meet the definition of ontological emergence" [12, p. 187]

On the other hand, epistemological emergentism may be defined as follow:

"A property of an object or system is epistemologically emergent if the property is reducible to or determined by the intrinsic properties of the ultimate constituents of the object or system, while at the same time it is very difficult for us to explain, predict or derive the property on the basis of the ultimate constituents. Epistemologically emergent properties are novel only at a level of description" [12, p. 186]

After this short outline of epistemological and ontological emergentism, in section 1.1 I shall discuss the ontological commitments of both kinds of emergentism, while in section 1.2 I shall analyze their dichotomy. Finally, in section 1.3 I shall

[1] European School of Molecular Medicine, Istituto Europeo di Oncologia, PhD student, FOLSATEC Group, IFOM-IEO Campus, Milan, Italy, emanuele.ratti@ieo.eu

show one case in which the dichotomy can be solved (that is, in the computer simulations of Artificial Life).

1.1 The ontological commitment of the emergentism as a mereological theory

I am not going to discuss here the informativeness and the coherence of both definitions outlined above. What I would like to discuss here is their ontological commitments. Considering that both contemporary epistemological and ontological emergentism try to solve the problem of the parts-whole relation in (complex) systems, they are arguably mereological theories. The ontological commitment of mereological theories is well studied by Achille Varzi [13] . Varzi puts the problem of the ontological commitment by saying that it is a matter of making 'an inventory of the world', i.e. a list of the things that exist. One of the main question linked to this task is whether we need to consider the parts of an object as things that "do not quite have a life of their own" [13, p. 283] and, more important, whether objects composed by parts should be considered as objects to be included in the inventory of the world. In order to solve this problem, we need what Varzi calls 'a count policy'. Emergentism is a type of count policy. Then Varzi continues by saying that there is minimal criterion that is able to state whether a count policy is plausible, but this not of interest here. What is important to highlight here, is that epistemological and ontological emergentism have two completely different count policies. They are not just different, they are opposite, i.e. they are irreconcilable. Let us see in details.

As I said above, ontological emergentism draws a layered-world imagery. At each level, there may be new objects (properties, phenomena), that are irreducible, unexplainable and unpredictable with respect to the lower level objects (properties, phenomena). What ontological emergentism tries to stress, is that there are sums of parts that are something (ontologically) new with respect to the parts they are composed of, so that we should not call these 'sums of parts', but rather 'wholes' because they are more than the sum of the parts they are composed of. We should count as existing in our inventory of the world every wholes (*composed by parts that can be wholes too*) at each level[1]. This is always true, until we meet the level of physics, in which the parts have no proper parts because this level is the last one in the layered model. We include also the basal objects without proper parts. Thus, even the consequences of this confused imagery are misleading, we can state that the count policy of the ontological emergentism (O) is the following one:

(O) The inventory of the world is to include the entities that are wholes with proper parts, and also the ultimate constituents of reality, that have no proper parts[2]

The count policy of the epistemological emergentism is different. The facts that epistemologically emergent properties are reducible to the ultimate constituents of the system, and that they are novel only at a level of description, give us the features of its ontological commitment. If here the wholes are nothing more than the sum of their parts, and the wholes are reducible to their parts, and the ultimate constituents of the world (i.e. what we should count as existing) have no proper parts, then the epistemological emergentism has the following count policy, call it E:

(E) The inventory of the world is to include the ultimate constituents of the world (whatever they are), so that no entity with proper parts should be count

Epistemological emergentism, in this counting principle and in Silberstein-McGeever's definition, is non-emergentist. The only reason why it is called 'emergentism' is because, in some sense, it explains the appearance of emergent stuff, i.e. there are some epistemological constrains so that we cannot reduce in practice the emergent phenomena to the basal ones even if in principle it is always possible. One may argue that epistemological emergentism can be also seen as completely uncommitted, in the sense that the arguing for more things to count (wholes, parts, etc) is not important, because for the purpose of the epistemological analysis of the epistemological emergentists, ontology is not relevant. Hence, you can argue in favour of epistemological emergentism without claiming that the supposed wholes that seem new are reducible to the entities of the basal level. In fact, you may merely make reference for the epistemological problems posed by the attempt to predict certain entities and their properties with respect to entities of the lower level. Bedau's weak emergentism [1] leaves open this possibility of an epistemological emergentism uncommitted ontologically. However, in this context this cannot be an objection, because I am focusing on those kinds of epistemological emergentism [12] [4] that have explicitly an ontological commitment. In the definition of epistemological emergentism I used, it is clearly highlighted that emergent entities and their properties are novel only at the level of description you are using, and the inventory of the world is committed ontologically to the ultimate constituents.

Although, at certain extent, the two inventories overlap (in the sense that their sets of things seem to intersect), the two counting policies are incompatible. Let us see why. O and E both accept in their policies the ultimate constituents that have no proper parts (namely, x). However, they do this for different reasons. O accepts x because its aim is to ground, in a quite physical and concrete sense, its levelism: it must admit a bottom-end to reality understood as hierarchy of levels. Thus, O admits x in order to ground the ontological hierarchy of those wholes with proper parts because, without x, there would be an infinite regression to lower and lower levels generating a confusing and endless inventory of the world. I am not able to make sense of such an infinite regression in ontology. On the other hand, E admits x for other reasons. E justifies the fact that emergent phenomena are novel only at a level of description, by arguing that there is one level that it is not only a level of description, but also a level of reality (the level of x). Thus, in order to say that emergent phenomena are only novel at a level of description, and that ontologically are not novel, it has to state that they are not novel with respect to something else, i.e. that they are reducible to x, and x should be considered as existing (so, included in the inventory of the world). In summary, E and O they are not compatible even if they both admit x, because O does this in order to ground the wholes with proper parts, and E in order to exclude the wholes with proper parts. Thus, they use x in order to ground two things that are not compatible.

I can provide an example of this incompatibility. For a taxonomist, the inventory of the world is composed mainly of

[1] Someone says that ontological emergentism is a rare phenomenon, someone does not, but it is not of interest here

[2] There is no space here to discuss which kind connection there should be between different parts. If the reader would like to go deeper in this problem, then he can have a look at the article of Achille Varzi cited above

organisms. Of course, if I ask him explicity to list his inventory of the world, he would include also such stuff as atoms (assuming for simplicity that atoms have no proper parts). However, he would do so only because the organisms he studies must be composed of some stuff, but he is not interest in atoms, rather in those particular features that can lead him to classify organisms in such and such a manner. Those features that he studies are considered ontologically new with respect to atoms and existing ontologically within the organisms (that they are new too) that bear them. A physicists will include in his list atoms, but not the organisms. He will say that organisms are entirely ontologically reducible to atoms, but of course the features of organisms, according to the context, may be considered at another level of description. Thus, they both include atoms in their inventories of the world, but for opposite reasons.

1.2 Metaphysical assumptions in the theory of emergentism

Ontological and epistemological emergentists hold, respectively, that O and E are not merely a way of modelling the system under scrutiny, but correspond directly to the structure of the system under analysis. Thus, it seems that reality is either O or E. This kind of dichotomy is classic in philosophy. For example, Floridi [3] analyzes a similar case, i.e. the dichotomy between the digital ontology vs. the analogue ontology. These kinds of dichotomies, viz. analogue vs digital or O vs E, are similar to the antinomies of Kant's pure reason , i.e. a thesis and an antithesis both reasonable and irreconcilable. Applying to these a trascendental methods [2], we shall show that the dichotomy E vs. O is misleading (of course when they both refer to the same objects), as Floridi shows for digital vs. analogue and Kant for the antinomies. So, we do not have to choose between O and E in this particular context. One case in which the 'metaphysical question' may have a reasonable answer is, as we will see, the case of a computer simulation considered as the system under scrutiny. But let us start with the trascendental method, i.e. the methods of levels of abstraction.

A trascendental method is a method that investigates the conditions of possibility of the experience (or, in this case, of modelling a system). Thus, we may analyze emergentism from this point of view in order to see whether we have some reason for stating that the inventory of the world of both epistemological and ontological emergentism corresponds directly to the structure of the world. Let us start with the notion of 'level of abstraction'. Floridi defines a level of abstraction as "a finite but non-empty set of observables, where an observable is just an interpreted typed variable, that is, a typed variable together with a statement of what feature of the system under consideration it stands for" , [3, p. 20]. Thus, a level of abstraction is composed by a particular set of features, that are supposed to belong to the system, that we take into consideration in order to observe the system. We commit ontologically to the system with these 'features'. Of course, the notion of level of abstraction is more complex than my simplification. For example, it is necessary to understand the relations that hold different observables, and there is also a way of putting together different levels of abstractions in order to make different kinds of observations. However, we want to show here only what happens when we conceive O and E as different ways of modelling a system, whether we have some reason to state that their manner of modelling the world corresponds to the structure of the world in itself. In order to do this, the notions of 'level of abstraction' and 'observable' are sufficient. A simple example may be the one about the differences between the analysis of a song made by a music reviewer and a sound engineer. The former shall look at the song through observables that are aesthetic categories, while the latter shall look at the song through the laws of physics. They look at the same system, but their analysis are completely different. So, the music reviewer, specifying that he is going to review a song from an aesthetic point of view, chooses a set of observables. In doing this, he (implicitly) specifies, as Floridi says in general for the levels of abstraction, "the range of questions that (a) can be meaningfully asked and (b) are answerable in principle" [3, p 22]. Thus, if we analyze the system only from a particular position, through particular observables, the information we can 'extract' from the system are limited, i.e. limited to the level of abstraction we use. This means that we cannot observe the system without specifying the position from which we are observing it, i.e. without specifying the level of abstraction we use.

Generally, in the debate of emergentism, it is assumed that reality in itself may be either O or E. This is a typical case of what Floridi define "the strive for something unconditioned" , [2, p. 317] that "is equivalent to the natural yet profoundly mistaken attempt to analyse a system (the world in itself, for Kant, but it could also be a more limited system) independently of any (specification of) the level of abstraction at which the analysis is being conducted, the questions are being posed and the answers are being offered" [2, p. 317]. This is because committing oneself ontologically to a system means modelling a system, so that we try to elaborate a model of the system. The only way to do this is to choose a level of abstraction, which determines the set of observables we use in order to elaborate the model. The ensuing model identifies a structure, which is attributed to the system. The problem is that the structure is identified only through a level of abstraction. We cannot avoid it. The emergentists do not take into consideration this problem, and even if they did, they implicitly assume that the observables of the level of abstraction they choose are the only available for observation in the system in itself, but we have no reason to believe that O or E correspond directly to the structure of the world (consider the example of the taxonomist and the physicists). We do have nothing in order to state that the level of abstraction we use has exactly the same features of the world in itself.

This is apparent if we consider that there is not a count policy that states which is the structure of the system we are analyzing, i.e. the structure of the system independent from any level of abstraction. The reason is that, and Varzi agrees in this, the count policy we choose depends on the underlying mereological theory, so that it is only a matter of choice. *In choosing the underlying mereological theory, we choose our level of abstraction*. The consequent count policy clarifies the set of observables we apply to the system. 'The strive for something unconditioned' in emergentism is misleading, because, if we cannot avoid the choice of a level of abstraction (the choice of the underlying mereological theory), then we cannot 'extract' something unconditioned from the system. The information we can extract from the system depends on the position we choose with respect to the system, i.e. the level of abstraction we choose.

1.3 Computer simulations and the noumenal

We saw before that the emergentism may take for granted a direct correspondance between its mereological theory (whatever it is) and the structure of the system under analysis. This is something it is impossible to prove, we have no direct access to what Kant calls the noumenal. In Kant's approach, only the creator of the noumenal may have access to it [5]. We shall show that the one case in which we can have access to the noumenal world is when we build the system under scrutiny (so when we build the noumenal), and this happens in the computer simulations, e.g. in the field of the artificial life (A-Life). It seems to be a trivial point, but it shows exactly in which sense there may be a direct correspondance between a level of abstraction and the noumenal world (that is what emergentists assume), and in which sense only the creator of the noumenal may have access to it.

Let us show this through an example. In [11] Ronald, Sipper and Capcarrère (RSC) try a different and experimental approach with respect to emergentism. The three researchers work in the field of A-Life, and they propose an emergence test, i.e. a sort of criteria to state whether one can use the label 'emergence' in different cases. The experiment works in the following way. There are two persons: a system designer (that designs a computer simulation through the A-Life), and a system observer. There are three stages of the experiment. The first (1) is the design, i.e. the system is built by the designer through "local elementary interactions between components (e.g., artificial creatures and elements of the environment) in a language L_1" [11, p. 229]. Then (2) the system observer observes the system, he is fully aware of the design, "but describes global behaviours and properties of the running system, over a period of time, using a language L_2" , [11, p. 229]. Finally (3) there is the surprise. This is the "the cognitive dissonance between the observer's mental image of the system's design stated at L_1 and his contemporaneous observation of the system's behaviour stated in L_2" [11, p. 229].

There is something ambiguous with (2). If the observer is fully aware of the design, so that he knows that the system features in themselves are such and such, then there is no reason why he has to adopt another kind of language or perspective. So, we can modify (2) stating that the observer is not aware of the system's design, but he becomes aware at (3), and here comes the surprise: after having observed the system's global behavior with L_2, he is surprised (becoming aware of L_1) of how the system is in itself.

The most interesting character here is the designer. RSC focus their attention on the observer, and on his surprise, but the designer for our purpose is more useful. In fact, when the designer designs a computer simulation (for example the emergence of a nest structure in a simulated wasp colony), the position from which he may observe the simulation after it is started, is a particular kind of level of abstraction. It is a particular level of abstraction because its observables are exactly the features of the simulation in itself. The designer of the simulation may have access to the 'noumenal' of the simulation because the simulation is built through the observables of the level of abstraction he decides to adopt. In the case of the emergence of a nest structure in a simulated wasp colony, we have a sort of operational epistemological emergentism because the designer tries to build the emergence, in this case the epistemological emergence. The point is that, in this example, E is not only the ontological commitment, but it is exactly the way by which the system is in itself.

2 CONCLUSION

In this article, after having clarified the nature and the ontological commitment of both epistemological and ontological emergentism, I have analyzed their assumptions that their metaphysical imageries correspond to the structure of the world. I have carried out this analysis with a transcendental method found in and , and I have reached the conclusion that the dichotomy E and O is conceptually messy. Finally, I have argued that the one case in which the dychotomy can be solved, is when we build the system under scrutinty (using A-Life), so that the observables of the level of abstraction adopted correspond to the structure of the system itself.

ACKNOWLEDGEMENTS

I would like to thank professor Luciano Floridi for suggestions on the first draft of this paper, the anonymous referee and professor Mark Bedau for their comments on the final version.

REFERENCES

[1] Bedau, Mark. 2008. Is Weak Emergentism Just in Mind? *Minds & Machines* (18):443-459.

[2] Floridi, Luciano. 2008. The Method of Levels of Abstraction. *Minds & Machines* (18):303-329.

[3] Floridi, Luciano. 2009. Against digital ontology. *Synthese* 168 (1).

[4] Hempel Carlo, Oppenheim, Paul. 1965. On the Idea of Emergence. *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*, pp 258-264. New York: Free Press

[5] Kant, Immanuel. 1998. *Critique of Pure reason*. Cambridge: Cambridge University Press.

[6] Kim, Jaegwon. 1993. *Supervenience and Mind. Selected Philosophical Essays*. Cambridge: Cambridge University Press

[7] Kim, Jaegwon. 1999. Making sense of emergence. *Philosophical Studies* 95 (1-2):3-36.

[8] Kim, Jaegwon. 2002. The layered model: Metaphysical considerations. *Philosophical Explorations* 5 (1):2 – 20.

[9] Kim, Jaegwon. 2006. Emergence: Core Ideas and Issues. *Synthese* (151):547-559.

[10] Morgan, C. Lloyd. 1923. *Emergent Evolution*. Vol. 34: Williams and Norgate.

[11] Ronald, et al. 1999. Design, Observation, Surprise! A Test of Emergence. *Artificial Life* 5:225-239.

[12] Silberstein, Michael, and J. McGeever. 1999. The search for ontological emergence. *Philosophical Quarterly* 50 (195):182-200.

[13] Varzi, Achille C. 2000. Mereological commitments. *Dialectica* 54 (4):283–305.