Gödelian Self-Reference in Agent-Oriented Software

BOLDUR E. BĂRBAT *, ANDREI MOICEANU**, IULIAN PAH ***
* Faculty of Sciences, "Lucian Blaga" Univ. of Sibiu, ROMANIA,

** Faculty of Automation and Computers, "Politehnica" Univ. of Timişoara, ROMANIA
*** Faculty of Sociology and Social Work, "Babeş-Bolyai" Univ. of Cluj-Napoca, ROMANIA

Consciousness is not an on/off phenomenon,
but admits of degrees, grades, shades.
DOUGLAS HOFSTADTER

Abstract: The paper aims at substantiating the first steps of a more general undertaking regarding self-awareness in agent-based systems, founded on Hofstadter's ideas and presented in previous papers (illustrating the broad-band technology potential from an anthropocentric and transdisciplinary perspective). This second paper focuses on computer science aspects, keeping a definite engineering perspective: the target is a generic architecture – based on Gödelian self-reference – for agent-based applications meant for open, heterogeneous, dynamic and uncertain environments. Specific objectives are: a) to defend the undertaking from an agent-oriented software engineering point of view; b) to detail the rationale for starting by Gödelian self-reference; c) to specify a stepwise approach affordable within a narrow scope; d) to present self-cloning as the basic software mechanism able to uphold this approach; e) to outline very roughly an experimental model. (Details and implementation issues regarding mechanisms and models are described in forthcoming papers.) Preliminary conclusion: the approach seems workable and the applicative potential seems significant.

Keywords. Agent self-awareness; Agent-oriented software engineering (AOSE); Open, heterogeneous, dynamic and uncertain environments (OHDUE); Gödelian self-reference (GSR); Agent self-cloning (ASC).

1 Introduction. The "Self-*ish" Meme

The general undertaking regards self-awareness in agent-based systems (ABS), is founded on Hofstadter's ideas [17], and was presented in [7] mainly to illustrate the broad-band technology potential from an anthropocentric (and transdisciplinary) perspective as well as in [9] in a larger interdisciplinary framework. Since the target is a generic architecture – based on Gödelian self-reference (GSR) – for agent-based applications meant for present-day environments (i.e., OHDUE), the key question is one of viability: is it suitable to consider self-awareness as relevant agent feature when many other – less abstract and elusive – strong agency characteristics are still regarded as luxury, even in current large-scale ABS? The simplified answer given in [0] was: system complexity makes it desirable, agent technology makes it possible, and approaching it by GSR (i.e., the agent clones itself – usually spawning a better architecture) makes it affordable. This answer must be substantiated, detailed, and refined in two directions: A) focusing on computer science aspects; B) keeping a definite engineering perspective. Hence, the specific objectives are: a) to defend the undertaking from an agent-oriented software engineering (AOSE) point of view [1]; b) to detail the rationale for starting by GSR; c) to specify a stepwise approach (workable within the narrow scope of a PhD [22]); d) to present ASC as the basic software mechanism able to uphold this approach; e) to outline very roughly an experimental model. (Details and implementation issues regarding mechanisms and experimental models are described in forthcoming papers.)

Within this framework, the Dawkins-like section title suggests four associations to the "selfish gene" [14]: a) most paradigms in contemporary artificial intelligence have an obvious memetic character. b) The memotype of "Self" shows a cognitive complexity similar to the structural complexity of the genotype. c) The "Self-*" memeplex invaded modern information technology (IT). d) Despite memetic likeness, the "duplicate me" instructions in the genetic code and the "I clone myself" message in this paper are fundamentally different. While the first two aspects set the paper's epistemic background, the last two are related to its very core.

As to the relevance of self-aware systems, the *AgentLink Roadmap* [1] is explicit: "Computational systems that are able to manage themselves have been part of the vision for computer science since the work of Charles Babbage. With the increasing complexity of advanced information technology systems, and the increasing reliance of modern society on these systems, attention in recent years has returned to this. Such systems have come to be called self-* systems and networks […] aspects of these systems include properties such as: self-awareness, self-organisation, self-configuration, self-management, self-diagnosis, self-correction, and self-repair". When the self-* functions are maximised – contrasted to direct user intervention – the approach is known as *autonomic computing* [18]. The rationale: unable to manage the system complexity involved by operating in OHDUE, humans transfer its cognitive component to the system. Hence, such systems must work more and more autonomously – like living beings, automata, or some recent software. As regards software, autonomous adaptive behaviour stems now mostly from combining biological and engineering mechanisms [13] and involves practically all weak agency characteristics [16].

Perhaps an even more relevant sign that the *self-** meme-complex – and above all its flagship "self-awareness" – is (re)gaining currently high consideration in IT was the *DARPA Workshop* [13]. Its *Report* [2] summarises: "The vision of a completely general-purpose theory and architecture for self-aware systems is certainly not yet the state of the art. It is, however, an excellent long-term vision in that it idealizes a strong thread of ongoing activity that is of both theoretical and practical interest. Machines do not need to be self-aware in the same way as humans do, but some forms of self-awareness seem to be useful. For example, the ability to determine what a system knows and does not know what it can do and cannot do, and how it can be driven over a period of time in a way that is consistent with its goals. Self-awareness can make the system more robust and self-repairing over a period of time."

The rest of the paper is organised as follows: Section 2 details the rationale for choosing GSR as starting point for agent self-awareness. Section 3 shows that a prudent approach to ABS meant for present-day environments (i.e., OHDUE) can be based on micro-continuity. That involves dedicated software mechanisms and successive prototyping, in this case, experimental models (Section 5). Initial conclusions (Section 6) seem promising.

As regards the language, for the sake of effectiveness it will be "convenient" (in the meaning given by Poincaré), i.e. anthropomorphous, in line with McCarthy [20], Dennett [15] and Anderson [3] (details are given in [7]). Hence, in this paper "awareness" and first of all "self-awareness" should be interpreted only in its metaphorical sense. (Nevertheless, at the horizon dawns the mesmeric, dubious, and risky meaning.)

2 Gödelian Self-Reference. The Missing Meme?

It is noteworthy that "self-reference" is not among the seven memes of the *Self-** memeplex mentioned in [1]. This absence is not surprising since self-reference is hardly considered a significant feature because it is: a) ordinary for recursive functions (in both mathematics and programming); b) implied by any kind of self-awareness (the very use of "I"); c) seen as a structural detail (e.g., a grammar form). Hence, to become acceptable as concept able to model Hofstadter's brainwave and, additionally, to simulate "strange loops" [17] in real world ABS – in short, to get memetic value – *self-reference* needs a qualifier. At this stage of the research, the less unsuitable label seems to be "*Gödelian*". (The main reason to choose it: it is the first entity of the "*Golden Braid*". Nevertheless, it is not the best choice since, in this context, GSR has nothing to do with first order logic, nor with Peano arithmetic.) In fact, it is the concept described in [17] reshaped in light of AOSE and of the target of this paper.

The profile of GSR, in the long journey from function call to meme, is set up in four steps:

- *Disambiguation*. From this angle it is appropriate to start with negative characterizations. Thus, GSR is not the self-reference in: a) Recursive function theory (as, for instance, in Kleene's fixed-point theorem). b) Recursive calls in programs (widespread since ALGOL-like languages became prevalent). c) Semiotics (e.g., in literary metafiction). d) Self-replication (e.g., self-reproductive systems copying themselves from industrial raw materials).

- *Working definitions*. *GSR* ::= kind of self-reference implicated in agent self-cloning. *ASC* ::= spawning an agent identical to its parent. An example at implementation level (for systems with Windows-like application programming interface (API); details in Section 4): the parent-agent main thread calls a "*CreateThread*" system function passing itself as parameter.

- *Features*. For the sake of conciseness, the features are asserted via contrasting them with other kinds of self-reference characteristics, from an (over)simplified pragmatic – here, engineering – perspective:

a) In GSR the entities linked by self-reference are *identical*, whereas in programming the callee – albeit self-similar to the caller – is less complex than the caller. That makes the fundamental difference: in recursive calls reducing complexity is necessary to solve the problem (avoiding infinite loops), while in contrast, for any "self" – natural or artificial, alike – it would mean deadly *involution* (for a sentence it means less words, for a drawing less lines, for a canon less sounds!). What self would accept such kind of "reverse life"? Hence, an agent has to refer to *itself*, not to a more and more simpler one.

b) On the other hand, since the agent is a process – now acknowledged as such by a formal standard [16] – the invariance refers to the "self" *per se* (i.e., as regards tha agent "I") not to its architectonics. Indeed, "When is One Thing Not Always the Same?" [17]: Due to its temporal dimension, an agent changes over time but evolution – as biology proves it – does not prevent self-representation. In this respect, GSR is not atemporal like the theorems it stems from, nor implying mere algorithmic sequentiality, nor requesting discrete time. In short, GSR involves entities acting (and varying) in time, not petrified in eternity. Moreover, such entities evolve "antientropically" like their living counterparts. (Thus, Escher-like drawings tend to become motion pictures and "strange loops" look rather like "strange whorls".)

c) At present, GSR involves only one conceptual echelon. However, "Jumping out of the System" [17] and an accordingly "meta"-perspective are not excluded for long-range developments.

d) ASC is closer to (self-)reproduction in biology than to (self-)replication in IT (since it is not based on redundant resources). ASC complies rather with the ancestor-progeny definition allowing "to distinct between the exact / inexact reproduction" [19].

- *Expectations*. First, those derived from [17]: such a nonconformist self-reference could be a matrix (in both its connotations: *medium* and *template*), for "strange loops", which in turn could lead to a stepwise emergence of (a primitive kind of) self-awareness, based on the fact – here maybe merely hope – that "Isomorphisms Induce Meaning" [17]. Then, based on a pioneer opinion: "Developing self-aware computer systems will be an interesting and challenging project. It seems to me that the human forms of self-awareness play an important role in humans achieving our goals and will also be important for advanced computer systems. […] Self-awareness is continuous with other forms of awareness. […] The forms in which self-awareness develops in babies and children are likely to be particularly suggestive for what we will want to build into computers" [21]. If these expectations should prove to be too great, at least GSR should provide a workable mechanism for improving agent architecture (as "Plan B" for real-world applications, mandatory to save the undertaking as applied research when the basic research target is too far).

3 Looping Towards Strange Loops

The overall approach was outlined in [7]. To impair redundancy, here are restated *in nuce* only aspects relevant to this paper, together with their implications regarding GSR via ASC but, to preserve consistency, the approach is described as a whole:

- Micro-continuity manifests itself at both the *conceptual* and the *implementation* level [4] [5]: the incremental nature of self-awareness (see motto [17]), allows starting with few features and going on stepwise. Likewise, enabling in this way generic architectures, both transdisciplinarity and affordability are boosted.

- Because of a large palette of restrictions (complexity, cost-effectiveness, hardware, logistics, research capacity and duration, etc.) self-awareness will be studied only for purely software entities. Passing from *robots* to *agents* may also reduce reluctance to interact with, since humans (both users and researchers) are more worried about (brute) force than about (primitive) intelligence.

- Though, to avoid undue agent behaviour, the owner should be able to enter a privileged interaction mode ("sic volo" speech acts). (Agent ethical behaviour is dealt with in [8] and [23]).

- Agents are real-time beings acting in OHDUE. The implications are powerful (e.g., the antientropic evolution mentioned in the previous section) and have far-reaching effects: a) Agent time (and the logic governing it) should be as close as possible to human time [6] [10] and ontologies have to mirror it (for instance, containing rules for active waiting or dynamic prioritising). b) Agents have to deal with uncertainty (e.g., making decisions based on abduction or even induction, as in e-Learning[1]). c) They must be highly reactive (most of them driven by environment stimuli). d) They must be proactive too (showing flexible initiative). Corollary: although some agent parts could be modelled algorithmically or objectually, the agent as a whole cannot be incarcerated in an "object-coffin" as common object-ori-

[1] That is one of the main reasons why e-Learning was chosen as the first real-world test field for self-referencing agents. However, the results are not elaborated upon, being presented in a related paper.

ented programming environments impose. That is why GSR refers to self-*cloning* (spawning a new dynamic, intentional software entity, an *agent*) not to merely self-*reproducing* (instantiating a new static, passive software entity, an *object*).

- Since agents must interact with humans in human ways, their ontology (in its original meaning, i.e., their sketchy "Weltanschauung") should comprise: *I* (software entity), *You* (master), and *Rest of the world* (context-relevant environment). Without involving (nor backing up) Smith's Knowledge Representation Hypothesis, here, for both theoretical and practical reasons, human-compatible knowledge representation is highly desirable. In short, propositional communication is almost unavoidable. (The reason is an unsure "author thesis" [10]: in interacting with interface agents, humans prefer symbolic communication but would like that their possible sub-symbolic response should be perceived too.) In addition, agents should be captologic and pathematic [5] [6] – especially for application domains where persuasion is vital (as in e-Learning). (Here micro-continuity can help since not the *antropomorphic feature* itself has to be replicated, but its *appearance* – firstly forged, later more genuine [5] [10].)

- Since this paper focuses on computer science aspects, the approach is refined considering the keystone *DARPA Workshop* [13] with its explicit conclusion: "In machines, self-awareness is likely to be of interest only for long-lived programs – programs that operate over a period of time, and potentially interact with the external world or other programs" [2]. That means *agents*. Although, because affordability is judged in line with other criteria, most basic ideas asserted or quoted there implicate a physical self-representation: "the self-representation of physical properties, known in virtue of an agent's perceptual connection with a particular object – his body", because "somatoception, the awareness of one's own body, involves many specialized sensors arranged into several distinct information systems" and "at this very basic level, self-representation is bodily-representation, and the self is known as, and in terms of, its body" [3]; "it follows from the simple fact that I somatically proprioceive particular bodily properties […] that those bodily […] properties are my own" [12]. That means *robots*. Luckily, there is hope for bodiless agents: "But naturally self-representation encompasses other kinds of properties besides the physical, among them intentional and self-reflexive. Intentional self-representation is, as the name implies, concerned with the ability to represent information about the intentional states of the self such as belief, desire and intention. Whereas at the level of physical self-representation the self is represented primarily as a body, at the intentional level the self is represented as an agent" [3]. The consequences are major (here are mentioned just two of them; details in future papers):

a) Lacking any spatial sense, the agent should excel with its temporal sense (after all, the user expects from an agent to react first in *real time* not in *real space*!). (This is another reason why agent time should be closer to human time.)

b) Unable to go or even to look "with free will", agents need a strong dose of "macroscopic non-determinism" to reach isomorphism to "macroscopic human free will" (this indefensible phrase is meant to avoid the "free will dispute"). Considering also "Plan B", OHDUE offers the most practical solution: algorithmic design should be drastically reduced according to the slogan "computing as interaction" [1] while wherever agent intelligence is really needed, object-orientation should be eliminated conceptually and lessened in implementation (a detailed rationale is given in [10]).

- GSR must be able to function also in the "Plan B" framework. As a result: a) ASC must be outlined as a mechanism with incrementally extendable functionality (see next section). b) Other mechanisms – preferably adapted versions of existing ones [4] [5] [6] [10] [11] – should be added to the development toolbox (outside the scope of this paper). c) To better the chances of applied incremental research, trends in AOSE – as well as in IT as a whole – should be considered thoroughly, avoiding any inertia in revisiting conventional software engineering (for instance, the role of algorithmic design mentioned above).

4 Self-referring Software, From Recursion to Self-Cloning. A Model

Since self-cloning (see Fig. 1) is proposed as basic mechanism, its two components must be clarified:

- *Cloning* is flexible. In this seeming oxymoron, "flexible" means that the differences between clones are initially kept minimal and may become extensive only after recurring cloning (a clone is just a "slightly altered alter ego" [4]). Thus, cloning is seen as a simple way to implement versatility and polymorphism – both vital features for a generic architecture.

- *Self-cloning* is conservative. In this seeming pleonasm, "conservative" means that the agent clones *itself*, preserving self-representation (its "I"), but not necessarily its old world model too. Instead, the model is regenerated through "phenotypical expansion", in short through elementary learning. Thus, recently acquired knowledge, brought in dy-

namically into the ontology, is transferred into the executable program representing statically the agent (i.e into the agent "genotype"). Of course, from the pragmatic perspective of an application, the process is seen rather as spawning "smarter progeny" (as shown for e-Learning, in a related paper).

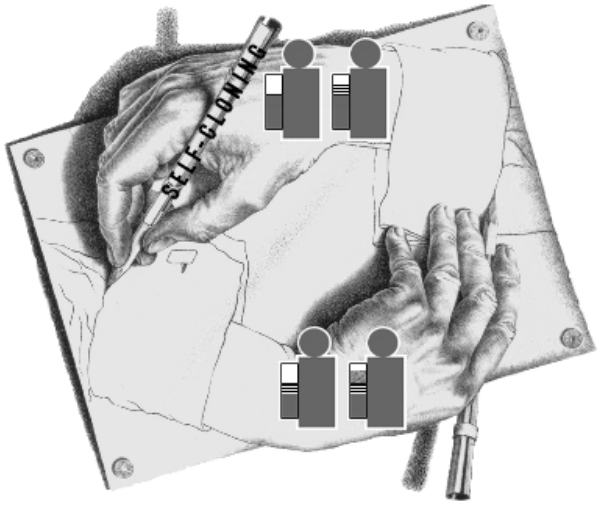

Fig. 1a. Self-cloning

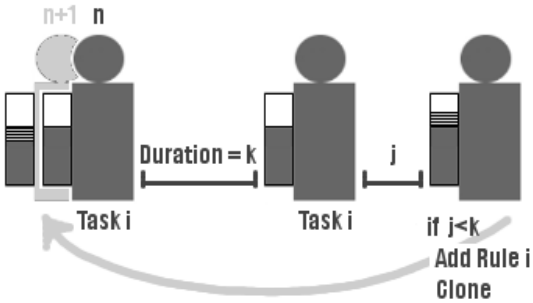

Fig. 1b. Learning between Clonings

Fig.1. Self-cloning: a "silicon copy" is not a "carbon copy"

Generic Achitecture. It is derived from a comprehensive design space, allowing to instantiate easily several applications that have to be: *diverse* (to allow conclusive incremental testing for both architectonic features, and mechanisms employed), *simple* (to adapt them fast for testing concepts and mechanisms, first separately, later merged), but intrinsic *usable* (to represent solutions first to toy problems but soon to real-world problems – albeit small-scale one – so as to be at least roughly conclusive) and easy *extendable* for further research.

Current Experimental Model. The present interface agent is carried out as pseudo-avatar, showing the following features: autonomy (adaptive intentionality, control of own mental states), longevity (it can leave a dying computer), basic sense of time (besides duration itself, context-driven waiting, dynamic prioritisation, etc.), effective self-reference (expressed by cloning itself after learning, hence improving its architecture), slave-interaction mode.

5 Conclusions: Rather *Desires* Than *Beliefs*. Anyhow, *Intentions*

1. From a computer science perspective it is much too soon to claim that agents could achieve self-awareness through Gödelian self-reference *per se*. Nevertheless, first indices are rather encouraging.

2. The main hindrance imposed by affordability restrictions is the purely software, bodiless, agent nature: the agent will lack the awareness of its own body, crucial for the somatoception-based self-representation achievable by robots.

3. Hence, the expected emergence of a primitive "I" should be catalysed through a powerful temporal dimension and an emphasised non-algorithmic behaviour.

4. Corollary: The main feature added to usual interface agent architecture and preserved through self-cloning is its primal sense of time. Besides its intrinsic architectonic value, it could be helpful in future "pseudosomatoception" as surrogate for the lacking sense of space and haptic proprioception.

5. On the other hand, "Plan B" is viable. In a Hofstadter manner of speaking, the agent could say: "I improve. Does it matter that I am yet uncertain about being (un)aware of it?"

6. The current agent endorses the model, and, mainly, the usefulness of self-cloning. In short, "Plan B".

As regards *intentions*, they are outlined in the stepwise approach: improve agent architecture, first of all its dynamic ontology, sense of time and reactivity (it should be much more event-driven).

Acknowledgement. This work was supported by the Ministry of Education and Research through Contract No. 73-CEEX-II-03/31.07.2006.

References:

[1] AgentLink III. *Agent based computing. AgentLink Roadmap: Overview and Consultation Report*. University of Southampton. http://www.agentlink. org/roadmap/al3rm.pdf, 2005.

[2] Amir, E., M.L. Anderson, V.K. Chaudhri. *Report on DARPA Workshop on Self-Aware Computer Systems*. Artificial Intelligence Center SRI International, 2004.

[3] Anderson, M.L., D.R Perlis. The roots of self-awareness. *Phenomenology and the Cognitive Sciences*, 4, 297–333, Springer, 2005.

[4] Bărbat, B.E. Holons, Agents, and Threads in Anthropocentric Systems. *Studies in Informatics and Control Journal*, **9**, 3, 253-268, 2000.

[5] Bărbat, B.E. Agent-Oriented Captology for Anthropocentric Systems. *Large Scale Systems: Theory and Applications 2001* (F.Gh. Filip, I. Dumitrache, S.S. Iliescu, Eds.), Elsevier, IFAC Publications, 214-219, 2001.

[6] Bărbat, B.E. Emotions and Time in Captological Agents. *Third International -NAISO Symposium on ENGINEERING OF INTELLIGENT SYSTEMS*, ICSC-NAISO Academic Press Canada/The Netherlands, 99 (Abstract; full paper on CD-ROM entityclosed), 2002.

[7] Bărbat, B.E., A. Moiceanu. I, Agent. *The good, the bad and the unexpected: The user and the future of information and communication technologies,* Institute of the Information Society, Moscow (forthcoming, May 2007).

[8] Bărbat, B.E., A. Moiceanu, H.G.B. Anghelescu. *Enabling Humans to Control the Ethical Behaviour of Their Virtual Peers*. Chapter in Enid Mante-Meijer, Leslie Haddon and Eugène Loos (Eds.) *The Social Dynamics of Information and Communication Technology*. (To be published by Ashgate, Aldershot, UK, 2007.)

[9] Bărbat, B.E., A. Moiceanu, I. Pah. Gödel and the "Self"ish Meme. *Gödel – Heritage and Challenge.* Interdisciplinary Symposium, Romanian Academy, Bucharest, 2007.

[10] Bărbat, B.E., S.C. Negulescu. From Algorithms to (Sub-)Symbolic Inferences in Multi-Agent Systems. *International Journal of Computers, Communications & Control*, **1**, 3, 5-12, 2006. (Paper selected from the *Proc. of ICCCC* 2006.)

[11] Bărbat, B.E., S.C. Negulescu, C.B. Zamfirescu. Human-Driven Stigmergic Control. Moving the Threshold. *Proc. of the 17th IMACS World Congress (Scientific Computation, Applied Mathematics and Simulation),* (N. Simonov, Ed.), e-book, ISBN 2- 915913-02-01, Paris, 2005.

[12] Bermúdez, J. L. *The Paradox of Self-Consciousness*. Cambridge, MA, MIT Press, 1998.

[13] DARPA. *Workshop on Self-Aware Computer Systems 2004. Statements of Position*. http://www. ihmc.us/users/phayes/DWSAS-statements.html#top

[14] Dawkins, R. *The Selfish Gene* (30th Anniversary edition). Oxford University Press, 2006.

[15] Dennett, D. *The Intentional Stance*. The MIT Press, Cambridge, MA, 1987.

FIPA TC Agent Management. *FIPA Agent Management Specification*. Standard SC00023K (2004/18/03). http://www.fipa.org/specs/fipa00023/ SC00023K.pdf

[17] Hofstadter, D.R. *GÖDEL, ESCHER, BACH: an Eternal Golden Braid*. (Including the Preface to the Twentieth-anniversary Edition.) Basic Books, New York, 1999.

Kephart, J.O., D.M. Chess. The Vision of Autonomic Computing. *IEEE Computer*, **36**, 1, 41-50, 2003.

[19] Luksha, P.O. Formal definition of self-reproductive systems. *Proceedings of the eighth international conference on Artificial life*, 414 – 417, MIT Press Cambridge, MA, 2002.

McCarthy, J. *Ascribing mental qualities to machines.* Technical Report, Stanford University AI Lab., Stanford, CA 94305, 1978.

[21] McCarthy, J. *Notes on Self-Awareness*. www-formal.stanford.edu/jmc/selfaware/selfaware.html, 2004.

[22] Moiceanu, A. *Self-Awareness in Agent-Based Systems* (PhD thesis in preparation.)

[23] Moiceanu, A., B.E. Bărbat. Ethical Behaviour of Self-Aware Agents. *The good, the bad and the unexpected: The user and the future of information and communication technologies*, Institute of the Information Society, Moscow (forthcoming, May 2007).