Artificial Agents and their Ontological Status

Denis BERTHIER

Professor (Artificial Intelligence and Epistemology)
Institut National des Télécommunications
Groupe des Ecoles des Télécommunications
9 rue Charles Fourier, 91011 Evry Cedex, France
phone: 33 1 60 76 41 22
mailto:Denis.Berthier@int-evry.fr
permanent URL: http://www.carva.org/denis.berthier

Abstract : Can an artificial agent "really" think? can it be "really" intelligent ? can it "really" have beliefs, goals, intentionality? Such questions have been plaguing AI since its inception. The solution we propose is to assert that an "artificial agent" is a *virtual* agent – and that all the cognitive or mentalistic attributes we may be tempted to grant it are also virtual. Whereas the first qualifier is purely descriptive, the second one is about the ontological status of such things as appear to us as agents. Our solution relies on: 1°) eliciting a precise core meaning for the word "virtual", a meaning that differs radically from the one used by philosophers (Bergson, Deleuze,…) but that has always been implicit in specialized contexts ("virtual image", "virtual world") and that should now be given its full ontological generality; 2°) relating the virtual to a broadened notion of interoperability, which justifies our assertion on technical and psychological instead of philosophical grounds. We relate this interpretation of AI to strong AI, to weak AI and to Dennett's intentional stance. We stress its implications for the cognitive sciences project of "naturalizing" intentionality. Finally, we mention some consequences in the information systems domain, relative to the acceptance of agent concepts for the modelling of organizations and their business processes.

Key Words : Epistemology, intentionality, Artificial Intelligence, virtual, interoperability.

I. INTRODUCTION TO THE PROBLEM AND TO THE PROPOSED SOLUTION

A question has been plaguing for decades all debates about Artificial Intelligence (AI), sometimes in so heated and inconclusive ways that it seems nearly everybody is sick with them and nothing new is to be expected. Can an "artificial agent" be "really" intelligent? can it "really" think? can it "really" have knowledge, beliefs, goals, intentions, emotions and so on? can it "really" have intentionality (in the full phenomenological, Husserlian sense), i.e. can it "really" refer to things in the world out there *and* to things in our heads? can it "really" communicate with us, share knowledge with us, collaborate with us – and what does all this mean?

Disregarding all the hype about AI and all the frustration it generated, the main dilemma initially raised by such questions is not only still present; it gets stronger than ever as AI goes deeper and deeper into the simulation of complex human behaviours:
– on the one hand, two very strong and complementary arguments, one technical and one psychological, plead in favor of an agent oriented vision : a) in order to design some complex

software systems, it is very helpful (or even necessary) to consider them as agents and to use some related formal mentalistic concepts[1]; b) moreover, in the proper context (i.e. in the operation framework it has been designed for), such an "artificial agent" is spontaneously perceived as having the aforementionned mentalistic attributes[2]; as a consequence one cannot be satisfied by just blaming AI (as is often the case) for a lack of precaution in its vocabulary;
– on the other hand, for philosophical, religious or common sense reasons, very few people are willing to "really" grant mentalistic attributes to a machine, however sophisticated it may be.

The solution we propose is to assert that an artificial agent is a *virtual* agent – and that all its cognitive or mentalistic attributes are also virtual. Here, the two qualifiers have completely different bearings: whereas the first one is purely definitional or descriptive, the second one is about its ontological status (where we understand "ontology" as being devoid of any form of essentialism). Moreover, if we decompose the expression, the adjective "artificial" means "produced by a technical activity", but the word "agent" does not suppose any specific definition of agenthood. So that *our assertion should be understood with the broadest scope: as soon as, in its limited operation context, a software component appears to us as if it was an agent (whatever our notion of agenthood may be, whatever the implementation techniques it relies on may be) and it ceases to do so outside this context, then it is a virtual agent; the same applies to any of the mentalistic attributes we may grant it.*

Our solution amounts to replacing a yes-or-no question by a question relating to ontological modalities in general. Admittedly, it is a nonsensical or a quasi void assertion and it is therefore a delusory solution, if one relies either on the standard meanings of the qualifier "virtual" (as they are recorded in the dictionnaries or as the word is used by most philosophers), or on its vague or nearly undefined meaning (as it is currently used on every occasion). Saying the agent is virtual may even not be completely new in this bare form[3]. However, what follows is new.

In the subsequent three sections, our assertion is first explained at three different levels of understanding: based on analogies, based on a new core definition of "virtual", based on the notion of interoperability. Then consequences are drawn regarding other interpretations of AI (strong AI, weak AI, Dennett's intentional stance) and the cognitive sciences project of naturalizing intentionality. The last section draws a consequence for information systems.

II. LEVEL ONE: ANALOGIES

[1] As to the practical scope of this remark, let us remind that, giving its full meaning to (Newell, 1982), sophisticated methodologies have been developed and are widely used in order to facilitate software development in terms of agents, such as Gaïa (Wooldridge & *al*., 2000) and KADS (Schreiber & *al*., 1993) or its multiagents version MAS-Common KADS (Iglesias & *al*., 1997).

[2] "Every human being is so much predisposed by naïve psychology to conceive his actions and those of other people as the result of their goals, intentions, desires and beliefs that *the least non human behaviour* is *irresistibly* understood as that of an agent equipped with an intention or a goal" (Jacob, 2004, p. 13 – translation and italics are ours).

[3] For instance, the FIPA (Foundation for Intelligent Physical Systems, the international association for the normalization of multi-agent systems) specifies in its norms that an agent has a virtual knowledge base; in this context, the word "virtual" may be understood with the meaning we have elicited, but FIPA has not noticed that such elicitation was necessary. We are not aware of any explicit claim about artificial agents as being virtual in any precise sense.

The example of a reflection in a mirror (which is technically a virtual image) is prototypical of the meaning we want to associate with "virtual". More generally, the following table (some terms of which will be explained later) lists the analogies we want to establish between the idea of an agent being virtual and the way this word is used in association with three other different phenomena. For each phenomenon, we display a short explanation and the framework defining its condition of possibility.

Phenomenon	Explanation	Operation framework
Virtual image	Propagation of light rays	Being in the proper light cone and limiting oneself to the visual modality
Virtual sound	Propagation of sound "rays"	Being in the proper spatial domain and limiting oneself to the audio modality
Virtual world	Sensorimotor interoperability	Being connected to the Virtual Reality apparatus and limiting oneself to the predefined modalities it supports
Virtual agent	Semiotico-cognitive interoperability	Being in the proper predefined communication situation (language, topic, comm. links)

There are four major points these analogies intend to stress.

First, a virtual image or a virtual sound or a virtual world is not a real image or a real sound or a real world, but it *is* in all cases plainly *actual* – as opposed to potential. The only restriction is that there are conditions for my effective perception of it (to see a reflection in a mirror, I must be in the proper light cone and look in the proper direction), but it is a fact that the satisfaction of these restrictions is extrinsic to the situation itself, and therefore not of a kind such that one could find anything potential in this situation. Given a situation A of an object in front of a mirror, my entering a situation where I perceive the reflection is an event that doesn't change anything about the mirror, the object or its reflection – neither in the way the event of a dam breaking could change the state of the water it held (and the flood status of the valley below it), nor in the way the event of a measure on a photon could change (i.e., in this case, determine) its polarization state. This event, which is undoubtedly relative to situation A, could be called possible or contingent on it, but, being totally external to situation A itself, it cannot be called potential (nor latent); nor can anything in situation A itself.

Second, in all of these examples, the notion of a "context of validity" or "operation framework" is inseparable from our notion of the virtual. Concerning an artificial agent, this must be related to the fact that it appears to us as an agent only in the proper communication situation; we are aware of no artificial system that had no severe limitations on its operating context – hype notwithstanding. Of course, there is much research effort to alleviate such limitations, but enlarged context does not mean unlimited context.

Third, in any of these examples, there is nothing subjective or imaginary (in the sense "folle du logis" often associated to the word "imagination").

Fourth, one could object that there is no predefined absolute reality and that anything could therefore be said virtual. But this would be absurd in all the previous examples, a virtual "object" being defined in opposition to its real counterpart. And there is no virtual world without a real aparatus in a real world to support it. We cannot say for certain what "real" means, but anyway virtual can be defined only in opposition to it.

III. LEVEL TWO: ELICITING A NEW CORE MEANING FOR "VIRTUAL"[4]

Considering etymology and the previous first three examples taken from science and technology, one can elicit *a new core meaning for "virtual": that which is not real but displays the full qualities of the real, in a plainly actual – i.e. not potential – way*. Contrary to the current meaning, *this definition distinguishes clearly the virtual from the potential* and *it allows to understand why the virtual can have real effects*.

The current standard meaning of "virtual"[5], inherited from medieval Scolastics (and from the invention it made of the pseudo Latin *virtualis*), entails "not in actual fact" and can therefore hardly be distinguished from "potential"; the word, with this meaning, has been used intensively by the French philosopher Bergson (in Bergson 1896/1970, but also in many other of his works); a kind of a theory of the virtual as a process was developed by another French philosopher, Gilles Deleuze, in "Difference and Repetition" (1968), claiming to formalize Bergson's notion (a claim that may certainly be debated); more recently, Gilles-Gaston Granger (1995) still adopts a similar meaning. Another current meaning of this word in colloquial usage is "nearly" or "quasi" or even "pseudo"; we consider it as mainly rhetorical, following a fashion associated with the information technologies boom.[6]

But both of these meanings are in total contradiction with the intended meaning in our previous examples, in expressions such as "virtual image" (in geometrical optics) or "virtual sound" (in the music or movies industries) or "virtual world" (in the "virtual reality" domain – VR). Although these expressions are recorded in dictionnaries, they have not yet been related to any general core meaning for virtual (and therefore they have no compelling philosophical implications).

On the basis of etymology, virtual is "what has the virtue of"; and virtue (from the latin *virtus*, itself derived from *vir* – man, hero) means quality with underlying strength. Therefore, following (Berthier 2004, 2005a, 2005c), we define "virtual" as "that which is not real but displays the full qualities of the real, in a plainly actual – i.e. not potential – way". It is then easy to check that the above examples satisfy this definition.

There is a major consequence: something virtual can have actual effects – for instance, one can cure agoraphoby by walking in the "open air" in a virtual world. Even if the world we are in at some moment is virtual, all the mental experiences and feelings we live in it are still fully real. In ours analogies, this can be compared to the fact that the ligth rays "issued from" a virtual image are real. Any definition of the virtual that does not distinguish it clearly from the potential makes it logically very difficult (even impossible) to explain this fact: how could something that is potential and remains unchanged have actual effects? This would logically amount to ask how the water in a dam, which has the potential to flood the valley below, could flood it without changing its own situation.

Let us now consider our assertion that an artificial agent is a virtual agent. This is the place to notice that, at a first level, it rested on a vague analogy: the artificial system appears

---

[4] Our work on the virtual was inspired by Wittgenstein's claim that "philosophy aims at the logical clarification of thoughts" (in his "Tractacus") and by its complementary idea that, in order to dispel our confusions, we should inquire how our "language games" are used (in his "Philosophical Investigations" §115): "A 'picture' held us captive. And we could not get outside it, for it lay in our language and language seemed to repeat it to us inexorably." Here, the "picture" was the idea that a simple yes-or-no answer was needed to questions raised about AI (intelligent or not? …).

[5] Trésor de la Langue Française, Encyclopædia Universalis, Wordnet…

[6] One should also mention a regional meaning, in the information technologies domain: "digital" – which seems to us unnecessarily restrictive.

to us as if it was an agent, in the way a virtual image or a virtual sound appears to us as if it was real; but there are contexts were these appearances vanish; therefore let us call them virtual. What the present definition of the virtual provides is a second level of understanding of the assertion, i.e.: a) a meaning more precise than this mere analogy, because it involves instead a general ontological modality, defined from a general phenomenological standpoint; *and* b) a first phenomenological justification based on this meaning: it is a virtual agent because one can observe that, in its operation framework, it displays the full qualities of an agent.

IV. LEVEL THREE: JUSTIFYING OUR ASSERTION ON THE BASIS OF INTEROPERABILITY

The next step provides a stronger justification for our assertion by explaining *why* the artificial agent appears to us as an agent – much as explaining (in terms of light rays) why a virtual image appears to us as a real image justifies calling it a virtual image. This is done by introducing a broadened notion of interoperabilty (which generalizes the technical meanings of this word in the information technologies domain) and eliciting its duality with our definition of the virtual.

This supposes an approach of AI that relies on its effective practices and results (Berthier, 2002, 2004) instead of on general claims of realizations to come "in the near future"[7]. And the effective results – the innumerable products that have reached industrial and commercial stage – are specialized agents that are designed according to precise methodologies to solve pre-specified types of problems in pre-specified operation frameworks (a fundamental notion, separating AI from science fiction[8]).

In this conception, *AI can be understood as aiming at developing semiotico-cognitive interoperability between Man and the computer (and VR as aiming at developing sensorimotor interoperability – so that both together aim at developing interoperability in the two major modalities of ordinary human experience)*. Moreover, the virtual is the fundamental ontological modality necessary for the natural description of phenomena or situations that can be explained in a more analytical, or more scientific, way in terms of interoperability – in the same way as virtual images are the phenomenological description of what could otherwise be described in terms of light rays and the laws of reflection and refraction. As a result, it is on a technical and psychological rather than philosophical basis that we can state the agent and its mentalistic attributes are virtual.

"Cognitive" interoperability (with quotation marks) between artificial agents is a purely technical notion; although it was not formulated in such terms, one can consider that "cognitive" interoperability was already the aim of the famous 1990 KSE (Knowledge Sharing Effort) project. To underline its roots in classical computer science, let us define it briefly, in a bottom up way, as the top of an ascending scale of abstraction levels:

---

[7] An example of such hype is Lenat's announcement of e-Cyc and its capacity for automatic knowledge acquisition from the Web (Austin Chronicle, Dec. 19, 1999). Under the title "Cyc Invades Cyberspace", he writes: "When e-Cyc becomes fully operational in early January, a thermonuclear explosion in the amount of information being pumped into Cyc's knowledge base is expected, with the result of Cyc becoming exponentially smarter and smarter". More than six years later, has anybody seen this invasion or this explosion? (This is not to deny the usefulness of CYC for practical purposes).

[8] One might object that this is a very restrictive conception of AI. But we have not yet seen any AI product that could be considered as displaying "general intelligence" independent of any restriction on its operating context (even learning always occurs in predefined conditions). The notion of an operation framework seems to be essential to AI.

- the lower levels of physical compatibility between computers and network equipments;
- the level(s) of network protocols and data exchange;
- the clasical level of data and programs interoperability: databases interoperability (with the two aspects of syntactic normalization – SQL – and "semantic integration" of databases schemas) and objects interoperability (through norms such as CORBA or Java RMI);
- the level of "cognitive" interoperability between artificial agents (through conformance to KSE or to the more recent FIPA specifications – which includes: normalized means for translating between knowledge representation languages, for referring to ontologies and for communicating by standardized messages).

The next step, semiotico-cognitive interoperability (without quotation marks) between Man and artificial agents, is a non obvious extension of the previous ones; it involves some form of communication between man and the machine, preferentially close to natural language; it may display different degrees. It means that, in its pre-defined semiotico-cognitive operation framework, the artificial agent appears to behave in the same way as a human agent would in the same situation and, in particular, that (to a predefined extent) some meanings *seem* to be shared between the user and the agent. Due to the restriction on the operation framework, this does not imply a positive answer to another sulfurous question: has the Turing test been passed? This means no more and no less than the agent has been designed properly, relative to its intended goals and operation framework as an agent. Semiotico-cognitive interoperability should therefore not be construed as a general *a priori* property of an AI system but as a *regulating goal* of AI systems development.

Interoperability is the technical notion that enables us to get rid of the identification paradigm (in which man and the machine are confused in some vicious circle) and invites us to think our relationship to the computer and to virtual agents rather than thinking ourselves as computers – as has long been the case in some AI or cognitive science circles (Berthier, 2005b).

V. WEAKER THAN WEAK AI, STRONGER THAN DENNETT'S "INTENTIONAL STANCE"

Differences with classical interpretations of AI can be stated: while strong AI is simply defused, our conception appears to be weaker than weak AI (or functionnalism) but stronger than Dennett's intentional stance.

According to our view of AI, strong AI is simply defused. Once a new ontological modality has been defined and it is shown appropriate to qualify the artificial agents of AI and their mentalistic attributes, asking if these are "real" is essentially pointless: would anyone care to ask if a virtual image is real? That would be like trying to lit a fire with a virtual image. The scope of this remark should not be underestimated, considering that, disguised in new clothes, strong AI ideas are still lurking behind many undue interpretations or overstatements of some cognitive sciences results (for instance in relationship with the ideas of symbol grounding or embodiement).

But our conception is also weaker than weak AI or functionnalism[9]: instead of considering all observable behaviours of an artificial agent (observable by any objective external means), we are only interested in such behaviours as are observable by a human subject in a situation of interoperability with the agent, in its predefined operation framework. This includes two complementary restrictions: one on the artificial agent and its limited operation framework; the other on the observer, which must be human and in the appropriate situation. In the same way as you can perceive a reflection in a mirror only if you are at the proper place, you can perceive intentionality (or intelligence or any mentalistic attribute) in an

[9] There are many conceptions of functionnalism (Pacherie,1993), but this is pointless here.

artificial agent only if you interoperate with it in the proper operation framework. And you don't mind if a fly does not perceive it.

Morever, in both situations, it is a fact that you have no choice: even if you know that what you are seeing is just a reflection, you cannot avoid seeing it; even if you know that what you are interacting with is just a software component, you cannot avoid understanding its behaviour in terms of folk psychology: beliefs, goals, intentions and so on; this is in full accordance with the way we have defined the virtual. This distinguishes us from Dennett's intentional stance (Dennett 1987). Dennett's vocabulary (words in italics afterwards) may suggest that we have a choice: the intentional *stance* is the *strategy* that consists in *interpreting* the agent's behaviour *as if* it was a rational agent whose actions are determined by beliefs and desires. This is undoubtedly very relevant from the point of view of system developement (which is not the one we are talking about here), but it is much too weak to describe the unsophisticated user's natural attitude.

VI. NATURALIZING INTENTIONALITY

Another consequence of our redefinition of the general modality of the virtual has been elicited in (Berthier, 2005c). It raises a new challenge for the cognitive sciences project of naturalizing intentionality[10]. Since we have shown that real and virtual intentionality must be distinguished, any "natural" explanation of intentionality should be able to make a distinction between these two modalities. But this means not only that it should explain intentionality on the basis of (biological, phylogenetical…) specificities of mankind, but also that it should do so in a way that cannot be applied to artifacts simulating these specificities. As a result, it is very difficult to imagine how general approaches, like Thom and Petitot's morphodynamics, could do the job. But it is no less difficult to imagine how approaches based on Varela's general idea of the embodiement of the mind could work by simply producing specialized levels of detail for this embodiement.

VII. INFORMATION SYSTEMS AND BUSINESS PROCESSES MODELLING

Practical consequences are not limited to AI. Consider the information systems (IS) domain. IS are a main factor of economic competition and they are tracking us in almost every part of our social lives. They have been rapidly changing in nature in the last decade: whereas they were traditionally considered as supports for the operational activities of an organization and they were essentially procedural, they now tend to be considered as supports for new kinds of activities (project management, decision making, innovation processes) that are highly interactive and very difficult to model in a procedural paradigm. Moreover, integrative IS (IS that can federate different organizations or different processes in one organization, through some kind of interoperability) and flexible or adaptive IS (IS that can adapt "easily" to the changing business processes in a rapidly evolving world) are becoming main themes for research.

Considering agents and associated concepts seems to be inescapable in the highly interactive and evolving IS of the future. At the technical level, multi agent systems have naturally been considered by many researchers as a possible answer to the above problems (Kishore & al., 2004; Wagner, 2003). But the same conclusion can be reached if we approach IS in terms that are meaningful for the organization it is designed to operate in and that allow one to make a link with its strategic orientations (Berthier & al., 2005).

---

[10] Here, we can consider any of the variants of this project (for a review of these variants, see Pacherie, 1993).

In fact, numerous differences can appear between an IS such as it was initially specified and designed and its final software implementation; moreover, after its introduction in an organization, numerous factors may have changed: in the organization, in its human actors, in its environment, in the development or maintenance team or in the system itself. So that, finally, the rationality that had initially led to the design of the IS can have become globally opaque to anybody in the organization. As a consequence, there is a risk that the IS tends to: a) locally, constrain every human actor by the interoperability modes it impones on him; b) globally carry a kind of autonomized rationality, becoming a kind of agent whose implicit micro-decisions elude every possible control – even if one does not want to consider it as such. With some exaggeration, it appears as if there was an army of ghost agents that secretely maintained alive past forgotten decisions of the designers. Explicitly introducing agents in the design would make all this more explicit and easier to change.

Nevertheless, and this is an instance of the dilemma we started with, the business world displays some blocking. The notion of an agent inevitably evokes a variety of other highly mentalistic notions linked to the characteristics of agenthood. Most practionners of IS are not likely to unconditionally admit introducing concepts such as an artificial agent, its goals and intentions, nor are they ready to hear about "social conventions" that would tie together such agents and real human agents. Stating clearly from the beginning that all such concepts are meant as virtual could make things much easier. Of course, we do not mean that this would solve the IS problems we started with. There remains a lot of work to be done; for instance: what ethic rules and legal regulations should apply to agent based IS? But having clarified the ontological status of the artificial agents should help tackling such questions.

VI. Conclusion

In order to solve the persistent dilemna of AI about the mentalistic attributes of artificial agents, this paper has introduced a new core meaning for the word "virtual", a meaning that distinguishes it clearly from "potential" – in opposition to all philosohical traditions but in accordance with many technical as yet isolated examples. As long as a software component appears to us as displaying some mentalistic attributes (intelligence, goal oriented behaviour, intentionality,…) in some operation framework, but can be denied these attributes in other contexts, then these attributes cannot be real; they have to be said virtual; and the agent itself must be said virtual. The new perspective on AI this definition allows has been related to its classical interpretations. Some practical consequences have been pointed out.

VII. References

Bergson, H. (1896/1970). *Matière et mémoire*. trad. fr. in Œuvres, PUF, Paris (also available on the Web).

Berthier D. (2002). *Le savoir et l'ordinateur*. L'Harmattan, Paris.

Berthier D. (2004). *Méditations sur le réel et le virtuel*. L'Harmattan, coll. « Impacts des Nouvelles Technologies », Paris.

Berthier D. (2005a). Virtuel / virtual, *DITL, International Dictionary of Literary Terms*, J.M. Grassin ed., http://www.ditl.info/arttest/art6573.php

Berthier D. (2005b). Penser notre relation à la Machine, plutôt que nous penser comme des Machines, *Terminal*, n° 92, pp. 155-170.

Berthier D. (2005c). L'intentionnalité et le virtuel, *Intellectica*, n° 40, pp. 91-108.

Berthier D., Morley C. & Maurice-Demourioux M. (2005). Enrichissement de la modélisation des processus métiers par le paradigme des systèmes multi agents, *Systèmes d'information et management*, Vol. 3 n° 10, pp. 25-45.

Deleuze G. (1968). *Différence et répétition*. Presses Universitaires de France, Paris.

Dennett D. (1987). *The intentional stance*. MIT Press, A Bradford Book, Cambridge, MA.

FIPA: http://www.fipa.org

Granger G-G. (1995). *Le probable, le possible et le virtuel*. Odile Jacob, Paris.

Husserl E. (1931/1994). *Méditations cartésiennes*. Armand Colin, PUF, Paris.

Iglesias C., Garijo M., Centeno-Gonzalez J. & Velasco J., Analysis and Design of Multiagent Systems Using MAS-Common KADS, *Proceedings of the 4th International Workshop on Intelligent Agents IV, Agent Theories, Architectures and Languages*, Lecture Notes in Computer Science, Vol. 1365, Springer, 1997.

Jacob P. (2004). *L'intentionnalité*. Odile Jacob, Paris.

Kishore R., Zhang H. & Ramesh R. (2004). Enterprise integration using the agent paradigm : foundations of multi-agent-based integrative business information systems, *Decison Support Systems*.

Newell A. (1982). The Knowledge Level, *Artificial Intelligence*, Vol 59, pp 87-127.

Pacherie E. (1993). *Naturaliser l'intentionnalité*. PUF, Paris.

Schreiber G., Wielinga B. & Breuker J. (1993). *KADS: A Principled Approach to Knowledge-Based System Development*, Academic Press.

Wagner G. (2003). The Agent-Object-Relationship Metamodel : towards a unified view of state and behavior, *Information Systems*, n° 28, pp. 475-504.

Wooldridge M., Jennings N. & Kinny D., The Gaia Methodology for Agent-Oriented Analysis and Design, *Autonomous Agents and Multi-Agent Systems*, vol. 3, n° 3, p. 285-312, 2000.