Identity: the Contribution of Contemporary Philosophy to Agent Design

Ines Di Loreto, Fabien Hervouet

▶ **To cite this version:**

Ines Di Loreto, Fabien Hervouet. Identity: the Contribution of Contemporary Philosophy to Agent Design. RR-12004, 2012. lirmm-00663359

HAL Id: lirmm-00663359

https://hal-lirmm.ccsd.cnrs.fr/lirmm-00663359v1

Submitted on 26 Jan 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Identity: the Contribution of Contemporary Philosophy to Agent Design

Ines Di Loreto
ines.diloreto@fastwebnet.it

Fabien Hervouet*
fabien.hervouet@lirmm.fr

Technical Report - January 2012

Abstract

This paper aims to present some perspectives about the agent design research community. Typically in perspective paper we try to understand what the trends are in order to foresee possible future directions. Exceptionnally, instead of looking at where the agent community is going analyzing numerical trends, with this article we want to suggest where agent research could go but is not yet going. During our analysis we discovered that a large part of the agent research community based its work on particular concepts derived from western analytic philosophy based on the logical approach. While applying a pure logic approach could be enough to model interactions for purely artificial agents, we believe that with the advent of mixed environments – i.e. environments involving both humans and agents – we have to focus on a more human-compatible agent design. For this reason in this paper we analyze the identity concept in both, agent design and philosophy focusing on uniqueness, mind, and body, in order to show that other paths could be studied in order to create different directions in agent design.

1 Introduction

This paper aims to present some perspectives about the agent design research community. Typically in perspective paper we try to understand what the trends are in order to foresee possible future directions. Exceptionnally, instead of looking at where the agent community is going analyzing numerical trends, with this article we want to suggest where agent research could go but is not yet going.

We are of the opinion that in the agent research community most of the current trends originate from the translation of particular concepts – mostly from analytic philosophy – which are only a particular western way to look

*LIRMM/CNRS, University of Montpellier, France

at philosophy and agents. With this article we want to suggest that other paths originating from philosophy could be taken into account in order to create different directions in agent design. It is obviously not the aim of this paper to exhaustively explore over 3000 years of philosophical research nor 60 years of agent design. For this reason the paper will focus only on 20th century philosophy and will only deal with the concept of identity in both agent design and philosophy. In fact while analyzing current trends we observed that while trying to model and reproduce humans and societies, agent design mostly does not use a structured construction of the identity concept.

This paper will not only show why a part of contemporary philosophy has already abandoned the dualism of mind/body but also that contemporary philosophy has an interesting way of considering the concept of uniqueness, performance and environment as strictly linked, interacting with each other. This philosophical analysis will clearly be of no interest to the agent research community if not linked to a parallel analysis of trends in agent design research. For this reason the core part of this paper is structured to compare the state of the art of the concept of uniqueness, body, and mind in the agent research community and in parallel in philosophy.

After the analysis conducted for this paper, we feel that a major part of the agent research community has fallen into the "utilitarian trap" by considering agents as artificially disconnected from our human world. We are of the opinion that applying a pure logic approach could be enough to model human-like interactions in a purely artificial agents environment. On the other hand, because of the advent of mixed environments – i.e. environments involving both humans and agents – we think we have to focus on a human-compatible agent design. This is mainly what motivated the double study – agent design / philosophy – we conduct in the rest of the paper.

2 The role of identity: from philosophy to agent design

In order to understand what importance the concept of identity could have in agent design we will start by analyzing the role of the concept of identity more generally in computer science. In the last few years – with the widespread popularity of social networks – the classical idea that suggests a perfect – but very simplified – match between identity and user name has been questioned. Today it is widely accepted that identities are negotiated based on the individuals' characteristics as well as on the community's state and aims [39]. From this point of view, the participants' positioning is based not only on individual moves, but on phenomena that are both context-shaped (e.g., the community/social network I decided to join) and

context-renewing (built throughout each interaction) [34]. The awareness of the importance of identity so well identified in psychology and social network analysis seems to have little impact on agent design, even if their stated purpose is to recreate some kind of human behavior. A possible criticism of the translation of this approach to agent design is that while any human individual's existence cannot be replaced by any other, and therefore does not fall into any general identity, in the multi-agent world an agent is a general concept – i.e. a generic piece of code – and can a priori be replaced by any other agent. While this abstraction is not false, this kind of uniformity can create problems when trying to model agents with human behavior interacting with human beings. Is it really true that the same starting point *should* have the same ending point passing through the same interactions? In order to understand the importance of this question we want to underline the already known fact that concepts are not neutral. For example, based on the way the two concepts of man and woman are conceived and linked to the sex equality concept they can lead towards two very particular societal systems.

On one hand Mary Wollstonecraft [40] – one of the founding feminist philosophers – theorized a society "controlled and organized in asexual forms" which does not take into account the membership of the citizen of one sex or the other. In this kind of society, there are no men and women, only "Universal Men" [1]. The cult of the universal is expressed through the reduction of sensible bodies and individual differences to the "One and Universal Man", also implying the equality of the sexes in the particular way seen above. In this kind of abstraction there are two negations: the negation of the *Body* (and thus differences in sex), and the negation of the *Individual*.

Wollstonecraft's point of view could be compared with that of Olympe de Gouges [20] who foresaw a society where men and women are not reduced to a single subject neutral and disembodied (precisely because disembodied is neutral). Women, she says in her "Declaration of Rights", take part in the founding of the state precisely because they are a "different subject".

It is evident that these two points of view imply a different use of the same general concepts which lead towards different societies.

Bringing the analysis back to agent design we believe that the oversimplification of the concept of identity is caused by the fact that computer scientists – and thus agent designers – based their work on the "Universal Agent" concept promoted by philosophy. The rest of this paper will analyze if and how the identity concept can be developed differently, paralleling agent design and contemporary philosophical assumptions about the con-

[1] Indeed, the genealogy of the concept of "Universal Man" should not be attributed to Mary Wollstonecraft but dates back to the Idea of Plato, Plotinus, and the republican ideals of the universal man so dear to the French Enlightenment.

cepts of uniqueness, body and mind, which are fundamentally negated by the use of the "Universal Man" concept.

2.1 The concept of uniqueness

As far as we know, uniqueness is a concept that has been studied in philosophy without entering the computer science sphere. This section presents possible reasons of this impermeability in the agent design community and explains recent philosophical theories computer scientists could try to integrate in their research.

2.1.1 The role of uniqueness in agent design

There is almost no debate about uniqueness in agent design. More generally uniqueness is in essence an issue for computer science. The transition from the analog era to the digital age has brought about the notion of copy. Copying is an intrinsic property of digital data. Any data can be copied and replicated with an absolute guarantee of ending up with two exact similar objects. This interesting property prevented researchers from really tackling such an issue.

Practically, in computer science, labels are the way we usually handle identity. In a computer language what we call variable is just a name referencing a specific memory space containing a piece of information. Data is stored in the memory and we can access it using suitable human-understandable names. Similarly, in multi-agent systems agents are uniquely handleable by their generated names, from the omniscient point of view.

Moreover, in MAS identity is mainly structured from the point of view of the role of the agents. Thus, MAS usually put a multitude of agents – identical for each role from the source code point of view – together in order to accomplish a certain global task or to have a certain global behavior. This means that even if agents do not act exactly the same in a local way, they often originate from the same piece of code that takes into account their interaction with their peers.

However, there is one research topic which addresses the severe lack of uniqueness in MAS: the trust problem. Indeed, in multi-agent systems each agent needs to be pretty sure of who it is dealing with especially if, for instance, it is dealing with economic decisions. In such a "hostile environment" the agent's goal is to choose its future interaction partners and shape these interactions based on its personal experience with the peers. The implemented techniques used in this case are the same used by human agents in online social communities, such as *reputation* (in the sense that all other agents may evaluate or devalue an agent based on its past interactions with him). See [32] for a general introduction on this issue and its associated challenges.

Another interesting point of view of the concept of identity/uniqueness applicable to computer science, comes from two biologists, Maturana and Varela, who addressed the complex problem of autonomy, knowledge and identity in [28]. In this work they characterized living organisms by coining the concept of *autopoiesis*. They defined autopoiesis as a complex process of self-production of the system by itself. Thus, an autopoieitic system should be seen as a machine that continuously generates and specifies its own organization. It accomplishes this incessant process of replacing its components because it is continuously subjected to external disturbances and constantly forced to compensate for these disturbances. In short an autopoietic system can be seen as a homeostatic system whose invariant principle is its own organization (seen as the network of relationships that defines it). Even if the authors don't use the term of *uniqueness* in their book, they clearly claim that identity is the product of the historical coupling between the organism and its environment. Therefore, uniqueness may be defined as this historical coupling, i.e. the historical adaptive activity of the structure in order to fit the organization. The autopoiesis theory has inspired some scholars in the artificial life and agent design domains such as [9, 7, 42], but it is generally very few addressed.

2.1.2 The role of uniqueness in philosophy

As we have seen from this short overview of the uniqueness concept, while seeing uniqueness as a problem most part of agent design tries to resolve it using a "labeled" approach (i.e. designing uniqueness only as a minor agent's attribute). As a result a certain part of agent design seems to work on a "Universal Agent" concept which does not need a structured uniqueness to be implemented. In addition the most interesting and structured point of view on uniqueness comes from "outside" the computer science domain. Similarly in philosophy there are different points of view regarding the concept of "Universal" man, and different criticisms which could help to open new paths in agent design. In particular feminist philosophers of the second wave did the more interesting work in destroying the predominance of the concept of "Universal" identity. Hannah Arendt before and Adriana Cavarero after, accused philosophy of constructing a universal "science" on the definition of "Man" declaring the uniqueness of each human being not "scientifically" defined and therefore unnecessary to his real existence in the world. Philosophy in fact always asks *what is* a supposed universal reality (the Man, the Being, the Subject, etc..) and ignores the other important question that living beings deal with one another: *who are you?*. In [1], Adriana Cavarero analyzes the important role played by Christine Battersby in defining the concept of identity. The philosopher focuses on a concept of personal identity consisting in a game of relationships that makes

the identity smooth, resonant and dynamic. For Battersby [6] identity it is not fixed and permanent as in the metaphysical tradition which loves to look at it as a "substance". It is not even the mere whirl of countless fragments as most of the contemporary feminist philosophy wishes it to be. Instead it is the identity of a relational self that persists without being the same, and becomes a fluid motion and permeable in its persistence. Apart from social, biological, and philosophical consideration the challenge of designing a self this way is radical because the self is thought of as *relational* – with other human beings but also with things – as well as a *singularity*, a uniqueness.

While uniqueness is a "new concept" in the long history of philosophy, psychology has discussed it from the very beginning of its history. In the '60s, Piaget led the constructivist movement, mainly in response to behavioral theories dominant at this time, considered as too simplistic. His vision is that every individual has the ability to hold their own reconstruction of reality. This educational philosophy has been radicalized by von Glasersfeld, who refutes the notion of external or ontological representation of reality. He introduces the *viability* concept, which is a relationship closer to the experience situated between knowledge and reality [38]. What he considers viable could be any action, a conceptual structure, a theory as long as it is useful to complete a certain task, to achieve a goal. He agrees with Piaget when saying that knowledge doesn't only aim to copy reality but it is more used for self adaptation.

This psychological movement, as a theory of knowledge, supports the fact that identity is perpetually in construction, deriving from our own adaptation process in direct confrontation with the environment. This point of view will be explored further in the '80s by Varela, Rosch and Thompson and their theory of *enaction*. In [37], the authors reintroduced the notion of one subject's unique experience as the center of the problem. The enaction paradigm postulates the co-emergence of both cognition and perceived world through the performative body in action in the environment. The authors also refuse any pre-given representation in favor of a complete construction of an embodied cognition. Each identity is then performative and unique, based on the subject's interactions. Therefore identity – and uniqueness with it – becomes a pure bottom-up mechanism that conflicts with the top-down approach.

2.2 The concept of body

The second aspect implied by the "Universal Man" concept is the abstraction from the body. As a matter of fact computers are not embodied, nor are computer programs. In this section we will see why and how the concept of body has been discarded by agent design researchers, because of the disembodied philosophical conception.

2.2.1 The role of the body in agent design

From the very beginning of computational science, the body hasn't been a priority at all. In essence calculus is disembodied as it is a pure mathematical abstraction, and computer science borrowed the same approach. However, regarding the "body concept" it is very interesting to notice a fundamental difference in the agent design approach between two domains: computer science and robotics. Indeed, while computer science focused mainly on disembodied reasoning capabilities, robotics was created with the idea of body, of physical interaction, and it is based on the experimental principles of physics and mechanics in a very grounded manner. Therefore, the advent of robotics is a major step towards the consideration of the notion of embodiment for intelligent agents, even if the transition from the automaton view to the agent view did not happen immediately in robotics, heavily influenced by the control engineering.

As Tom Ziemke argues in [42], it has been recognized for a decade that embodiment is a necessary condition to characterize living organisms. More and more researchers have attempted to address this absolute need for embodied cognition (see for example [31, 37, 17]). Ziemke also says that despite a general acknowledgement, robotics "is largely 'stuck' in the old distinction between hardware and software". He clearly claims that Searle's famous Chinese Room Argument (CRA) [35] – which says in short that disembodied manipulation of symbols cannot be considered as an authentic and intelligent understanding – remains unsolved. The main reason for this claim [41], is that the cognitivism vision – seeing the mind as a computational tool, i.e. manipulating symbols and rules, completely disconnected from any physical reality – is still dominant over the enaction vision.

However a kind of reactualization of the CRA has been proposed by Stevan Harnad [25] and its Symbol Grounding Problem (SGP). The SGP questioned the research community in a more precise way: will an artificial agent one day be able to develop a semantic autonomous capability allowing it to establish its own semiotic networks connecting some of its own symbols to the environment it evolves in? A lot of researchers have worked on this problem, especially Luc Steels who was a pioneer with his *Talking Heads* experiment [36].

Nevertheless the real revolution in robotics has occurred mainly within the last decade, during which we have seen the rise of the new discipline of developmental robotics. This field lies at the intersection of a number of scientific and engineering disciplines including at least developmental psychology, cognitive science, artificial intelligence, robotics, and philosophy. The key idea of developmental robotics is "physical embodiment" which considers the body as something essential in the realization of cognition and action. Developmental robotics explains that the body specifies the con-

straints on the interaction between the agent and its environment. Thus, research is axed on the sensorimotor mapping at the early development stage, and on the following stage of social development, aiming to fill the gap between these two phases in human life. See [5] for a recent survey of this quite young but very promising research domain.

2.2.2 The role of the body in philosophy

As we did for the uniqueness concept in the previous section, in this section we explore the idea of body from a philosophical point of view to see if we can add some interesting concepts to our computer scientist point of view. The first concept we will analyze is the one of "constructed body". Judith Butler [15, 16] states – on the wave of Simone de Beauvoir but also Monique Wittig – that not only gender is socially constructed but also the body is socially constructed. What the author means with this statement is that not only can a body be physically manipulated but also that gender, sex and body prove to be *performative*, capable of building identity through performance. For the above mentioned authors, identity is constructed through both a performative mental approach and a performative bodily approach. At the core of this statement there is the request to abandon the Cartesian (but also Platonic, Kantian and the like) mind / body dualism. In addition, the body is the instrument of our contact with the world, namely the world takes on a different aspect according to our way to grasp it. The body-object described by scientists exists only in the concrete experience from the subject, and is not an abstract entity. However, because of this idea the "body" often appears to be a passive medium that is signified by an inscription from a cultural source figured as "external" to that body. It is exactly this kind of separation that agent design borrowed from philosophy.

But what does having a performative body mean? In Judith Butler's article, "Performative Acts and Gender Constitution: An Essay in Phenomenology and Feminist Theory," the author argues that philosophers rarely think about acting (in the theatrical sense of the term) – this is more the habit of psychologists – but they do have a discourse of "acts" that maintains associative semantic meanings with theories of performance and acting. For example, John Searle's 'speech acts' seem not only to refer to a speaking relationship but to constitute a moral bond between speakers. Further, 'action theory', a domain of moral philosophy, seeks to understand what it is 'to do' prior to any claim of what one ought to do. Finally, the phenomenological theory of 'acts', by Edmund Husserl, Maurice Merleau-Ponty and George Herbert Mead among others, seeks to explain the mundane way in which social agents constitute social reality through language, *gesture*, and all manner of symbolic social sign [14]. However, even in this case, they seem to assume the existence of an agent existing prior to performances and

language.

When Simone de Beauvoir claims, "one is not born, but, rather, becomes a woman," [19] she is appropriating and reinterpreting this doctrine of constituting acts from the phenomenological tradition: identity is created through the subject performances/acts. However for her there is no stable a priori identity from which various acts proceed; rather, it is an identity tenuously constituted in time – an identity instituted through a stylized *repetition of acts*. A dynamic identity where the body is both, subject and object of the construction. Not only is identity both body and mind, but like for the uniqueness concept it is dynamically constructed in an environment. Once again we are facing an approach pushing towards a bottom-up identity contruction. The presence of a human being in the world strictly implies the presence of a body that is both *a thing of the world* and a *point of view over the world*. Donna Haraway [24] will push this interpretation to claim for a condensed image of both imagination and material reality – the two joined centers structuring any possibility of historical transformation – achieved through the cyborg's image. In addition the body is relational. The question *who are you?* that we mentioned before then takes on a special connotation. Every human being is unique because by simply exposing himself to the gaze of others, he is already unique in the body shape and tone of voice [1], *even before his name is known*. That is even before the label we attribute to him. Existence is therefore this mutual exposure *in a shared space* where everyone, from birth is *unique*, and in the course of his life can show who he is with acts and words.

In the previous section, uniqueness was defined as a historical coupling, i.e. the historical adaptive activity of the structure in order to fit the organization. In this section we are proposing a similar definition based on the interaction body/environment. While computer science looks at the body as quite a stable "hardware" element, philosophers suggest that the body is both a complex cultural construct, and a physical object interacting with the world. The body then is not a tool through which the mind manipulates the world, and it is not an attribute of the mind. On the contrary it is a historical *adaptive* coupling.

2.3 The concept of mind

As previously said, if the "Universal Man" is a disembodied entity then the mind and its thoughts become the central part of philosophical and scientific reasoning. This section is devoted to analyze this kind of approach and its criticism.

2.3.1 The role of the mind in agent design

From the historical point of view, we can easily claim that agent design started with the birth of cognitive sciences with the Macy's conferences in 1946. We can also bring back at this time the birth of the conceptual gap of how to consider human cognition – represented by the two famous researchers John von Neumann and Norbert Wiener – even before the construction of the first computers. On one hand, von Neumann considered cognition as fundamentally oriented toward problem solving, while Wiener claimed that cognition is more like an autonomous and self-creative activity, which he thought more credible to describe living systems. von Neumann's vision has been widely promoted and become dominant, giving birth to computer and engineering sciences. Besides this, computer is also the dominant metaphor used to design the brain, which is mainly based on the idea of an information process from inputs to outputs [43]. The other idea of autonomous aspects, coming from living systems, has been almost completely neglected. According to [28], von Neumann advocates the heteronomous systems, while Wiener and his cybernetics advocates the autonomous systems. In other words, heteronomous systems think in terms of I / O, while autonomous systems think in terms of operational closure. Always according to [28] while heteronomous systems are in a representational relationship with the environment, autonomous systems are in a meaning-emergent relationship with the environment. This implies two different logics: the logic of correspondence for heteronomous systems and a logic of consistency for autonomous systems.

For a very long time agent design has only taken into account the omniscient point of view. Indeed at the beginning, researchers thought it was a great opportunity for the computer to know exactly the whole system data. Thus they applied this omniscient strength to their algorithms. As an example, just think about how the $A^\star$ algorithm tries to solve the problem of moving from one point to another for a situated agent through graph theory, from a pure omniscient point of view. This kind of approach is quite similar to the philosophical one we talked about in the previous section, and which perceives the agent as en *external abstract entity*. While the algorithm is effective, it is a good example of the predominance of the mind over the body, even when talking about issues like displacement involving the body directly.

Furthermore, for a large majority of multi-agent frameworks, the environment is only seen as an opportunity of communication between agents. Besides, in agent design it is very interesting to inspect the way we make agents communicate. Communication is a root mechanism in computer science – just think about the case of message passing in the object oriented paradigm. Thus there was then no reason not to also provide soft-

ware agents with a communication mechanism. Thus in the '90s two main standards emerged in multi-agent systems in order to formalize communication: *KQML* and *FIPA-ACL*. These standards are philosophically based on Austin's and Searle's theories about speech acts, profoundly rooted in a pure mind-only logic. This is another example of the predominance of the mindfully dominant vision in the artificial intelligence community.

Nowadays, although the metaphor of the agent as a symbol interpreter is always present, things became more complex. For instance, Russell and Norvig's proposed a more complex approach to this issue. In their modern approach to artificial intelligence [33] – which is considered as a major reference in the AI teaching and research community – they distinguish four kinds of agents: reactive agent (which selects an action from its current situation and a set of rules), model-based agent (which extends the reactive agent by having a memory), goal-based agent (which tries to reach a state usually using planning), and utility-agent (which tries to maximize its satisfaction by trying to reach the most satisfying state). In the same movement we can also cite the Belief-Desire-Intention model [10] which is a more complex model articulated around the notion of knowledge: beliefs represent the information (true or false) the agent has about the world, desires represent long term goals the agent would like to accomplish, and intentions represent short term goals according to the desires.

In 1986, Minsky – after long years of collaboration with Papert – proposed a different perspective in his *Society of Mind* [29]. In this book he presented a cognitive model in which the mind is considered as a distrubuted multi-agent system with local and global adaptation. These agents are organized into agencies which can also be organized in network of agencies. His approach was very innovative and considered by Varela as a *middle path* of cognitive science between centralized cognitivism and distributed auto-organization [37]. However, even in these cases we faced with mind-centered approaches.

As we discussed in the previous section, the importance of the body was put back into the ring thanks to robotics. But robots are just embodied agent who also need to be designed with a mind. This is the reason why the way robots interact with their environment has been thoroughly studied for fifty years. Three major kinds of robotic architectures emerged. Deliberative architectures were historically the first ones to be proposed [30, 23]. These architectures use symbolism and are generally organized into multiple hierarchical layers, each layer communicating only with its direct superior and/or inferior neighbor layer. But these deliberative architectures lack responsiveness. In direct opposition, purely reactive architectures were considered [11, 2, 4]. They are built by stacking finite state machines called behaviors, directly connecting sensors to actuators in a completely reactive

way. They focus on the idea that complex behavior can emerge from a composition of simple reactive behaviors, without reasoning nor symbolization. Moreover, a few years later Brooks released two major articles [13, 12]. The first one was an accusation against the notion of representation on which AI is based. In the second one he put forward the many benefits of bio-inspired architectures. Subsequently hybrid layered architectures [21, 22, 8, 3, 27] have been developed in order to combine the advantages of behavioral and deliberative architectures. Thus they have both a hierarchy of layers dedicated to symbolization and decision making as well as reactive nested loops allowing each layer to provide appropriate responses to the dynamics and the operational reality of the robot.

2.3.2 The role of the mind in philosophy

In the 20th century problems deriving from the division between mind and body also become evident in philosophy. One of the most well-known – and consciously or unconsciously used for designing agents – 20th century philosophical movements is analytical philosophy. Very simply, analytical philosophy is characterized by the application of a logical method to traditional philosophical problems often using modern formal logic and language analysis[2]. After the 1960s such a "pure" logical approach has been questioned and now involves a much more general notion of an "analytic" style, characterized by precision and thoroughness of a narrow topic and opposed to informal and imprecise discussions of broad topics. However, even this approach seems to create debates among analytic philosophers. In a very recent article [18] Tim Crane – one of the most important contemporary analytic philosophy philosophers – offers his interesting perspective on the role of analytic philosophy in recent years. In particular in his article he focuses on the study of the mind. Mind in fact has always been the center of interest of analytic philosophy since the work of Bertrand Russell. To study the phenomenon of belief, for example, analytic philosophers looked at the ways in which we *talk* about beliefs, as well as logic and other properties of our discussion. But there is another reason why analytic philosophers have approached the study of mind and language and it arises from the influence of Wittgenstein on one hand, and by behaviorist psychology on the other.

As we know Wittgenstein argued the impossibility of a "private language", or of a language that was understandable by just one person. If we want to express our thoughts through language, what we express must be publicly accessible in some way: thought in its essence cannot be private. The behaviorist psychology of Skinner, starting from different assumptions,

[2]Some of the most important philosophers of this tradition are Bertrand Russell, Gottlob Frege, Ludwig Wittgenstein, Willard Van Orman Quine, Saul Kripke and David Lewis.

leads to similar conclusions: the only way we can study thought is to look at verbal behavior because, unlike in private thoughts, the behavior can be *scientifically* verified. The legacy of these perspectives in analytic philosophy is the vision of the mental representation (what Franz Brentano in the nineteenth century called "intentionality") heavily based on the study of language [18]. Crane continues asserting that while the approach to mental representation intent – based on language – has given many important results it also has a distorting effect on the correct understanding of the conscious mind.

The German philosopher Edmund Husserl – founder of phenomenology – saw intentionality and consciousness as intrinsically linked: the intent is an intuitive idea, it is how we imagine the world. In addition much in our representation of the world, in perception, thought and emotion, is unconscious. The approach to mental representation based on language is then unable to account adequately for consciousness. While Crane moves towards a doctrine where consciousness is the study of ideas in the private and subjective mind (fundamentally psychology-based) we want to argue for a more extended approach. As we can see, even in analytic philosophy there is a criticism leading towards the concept of subjective mind and uniqueness. While analytic philosophy seems to understand problems arising from the separation of mind and body it does not address the uniqueness problem, nor the body integration in depth. If we consider most of the logical approaches derived from the above mentioned philosophers then it is not surprising that agents have been conceived in the way described before. With regard to the language, for example, what agents designers did was closer to creating a communication language between agents – the explicit part inside thinking – rather than a language for a thinking being.

3 Perspectives

At the end of this paper we can then say that our provocative title can be explained this way: in the present state of things agents cannot have a sex not only because they have an abstract body – manipulated by an abstract mind – but also because "Universal Man" has no sex. In fact this paper started with the consideration that in the agent research community most current concepts regarding identity originate from the translation of a particular philosophical concept, the "Universal Man". For this reason the paper has shown that a part of contemporary philosophy and agent design has already abandoned the dualism of mind / body. In addition contemporary philosophy has an interesting way of looking at the concept of uniqueness, performance and environment as interlaced and interacting. At the same time this paper has underlined the fact that the approaches

based on logic and symbolic representations have the possibility to shift towards more complex approaches. This could be done using more recent philosophical concepts like, for instance, the ones we analyzed in this paper at least from a theoretical point of view. Although the problem solving vision is useful in many ways, integrating different concepts can lead to a more global vision about autonomous agent design.

Obviously identity is only one of the concepts that could be analyzed and the analysis proposed in the core part of this paper makes up only a subset of the concepts that can consist in identity. For example a whole section could be dedicated to the relational self – i.e. the way in which an agent holds a representation of other agents – vs. collective self – i.e. the way a group of agents create their identity. However we believe that even through this brief analysis we can underline interesting paths agent design can follow. Evidently the choice of which kind of concept to use is not neutral and would influence the final system. At the same time we are aware that agent design is an action-oriented field which needs to avoid the trap of "analysis paralysis" and to move towards knowledge scaling. For this reason we suggest replacing a top-down logical approach with a bottom-up enactive approach. Based on our analysis we suggest that agent design can integrate the following concepts.

Uniqueness The concept of uniqueness could be very interesting to integrate in agent design for example in a mixed environment – involving virtual agents as well as human agents. In this case uniqueness can be used by human agents for connecting with their own virtual ones.

Autopoiesis The concept of autopoiesis – maintaining organization through the evolution of the structure despite the environmental disturbances – is strictly linked with the uniqueness one. In autopoietic systems uniqueness may be considered as a particular trajectory of the coupling between organization and structure.

Enaction The concept of enaction could be integrated in agent design in order to overcome the dualism of mind / body. Going beyond this dualism can help to create agents which are more adaptive to unknown environments because of the physical grounding of their interactions.

The concepts we suggest integrating in the agent design paradigm are nothing more than necessary steps – but not necessarily sufficient – to reach the autonomy stage. However we strongly believe that as long as the design of agents is mainly based on analytic philosophy, we can only have an enlargement of the domain and not a paradigm shift – i.e. a change in the basic assumptions, or paradigms [26] – which is at the basis of science.

References

[1] F. R. Adriana Cavarero. *Le filosofie femministe*. Paravia Scriptorium, 1999.

[2] P. Agre and D. Chapman. Pengi: An implementation of a theory of activity. In *Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI-87),*, pages 268–272, 1987.

[3] R. Alami, R. Chatila, S. Fleury, M. Ghallab, and F. Ingrand. An architecture for autonomy. *The International Journal of Robotics Research*, 17:315–337, 1998.

[4] R. C. Arkin. Motor schema based navigation for a mobile robot: An approach to programming by behavior. In *International Conference on Robotics and Automation*, volume 4, pages 264–271. IEEE, 1987.

[5] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida. Cognitive developmental robotics: a survey. *IEEE Transactions on Autonomous Mental Development*, 1:12–34, 2009.

[6] C. Battersby. *The Phenomenal Woman: Feminist Metaphysics and the Patterns of Identity*. Routledge, New York, 1998.

[7] R. D. Beer. Autopoiesis and cognition in the game of life. *Artificial Life*, 10:309–326, 2004.

[8] R. P. Bonasso, R. J. Firby, E. Gat, D. Kortenkamp, D. Miller, and M. Slack. Experiences with an architecture for intelligent, reactive agents. *Journal of Experimental and Theoretical Artificial Intelligence*, 1997.

[9] P. Bourgine and J. Stewart. Autopoiesis and cognition. *Artificial Life*, 10:327–345, 2004.

[10] M. Bratman. *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA, 1987.

[11] R. Brooks. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2(1):14–23, 1986.

[12] R. Brooks. Intelligence without reason. *Computers and Thought - IJCAI'91*, pages 569–595, Jun 1991.

[13] R. Brooks. Intelligence without representation. *Artificial Intelligence*, pages 139–159, Nov 1991.

[14] J. Butler. Performative acts and gender constitution: An essay in phenomenology and feminist theory. *Theatre Journal*, 40(4):519–531, 1988.

[15] J. Butler. *Bodies That Matter: On the Discursive Limits of Sex*. Routledge, 1993.

[16] J. Butler. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, 1999.

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

[18] T. Crane. Con le migliori 'intenzioni', 2011. Retrieved September 13, 2011 from http://www.ilsole24ore.com/art/cultura/2011-08-28/migliori-intenzioni-081641_PRN.shtml.

[19] S. de Beauvoir. *The second sex.Translated from the French and edited by H. M. Parshley*. New English Library, London, 1970. Originally published by Libraries Gallimard in 1949 as Le deuxieme sexe.

[20] O. de Gouges. *La musa barbara. Scritti politici (1788-1793)*. Medusa, 2009. Collection of writings originally published between 1788-1793.

[21] R. J. Firby. *Adaptive Execution in Complex Dynamic Worlds*. PhD thesis, Yale University, 1989.

[22] E. Gat. Integrating planning and reacting in a heterogeneous asynchronous architecture for controlling real-world mobile robots. In *AAAI*, pages 809–815, 1992.

[23] G. Giralt, R. Chantilia, and M. Vaisset. *An integrated navigation and motion control system for autonomous multisensory mobile robots*. Springer-Verlag New York, Inc., 1990.

[24] D. J. Haraway. *A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century*. Free Association Books / Routledge, 1991.

[25] S. Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42:335–346, 1990.

[26] T. S. Kuhn. *The Structure of Scientific Revolutions*. University Of Chicago Press, 1996.

[27] D. Luzeaux and A. Dalgalarrondo. Harpic, an hybrid architecture based on representations, perception, and intelligent control: A way to provide autonomy to robots. In *International Conference on Computational Science*, pages 327–336, 2001.

[28] H. R. Maturana and F. J. Varela. *Autopoiesis and Cognition: The Realization of the Living*. Reidel, 1980.

[29] M. Minsky. *The society of Mind*. Simon & Schuster, 1986.

[30] N. J. Nilsson. *Principles of Artificial Intelligence*. Tioga Publishing Company, 1980.

[31] R. Pfeifer and C. Scheier. *Understanding intelligence*. MIT Press, 1999.

[32] S. D. Ramchurn, D. Huynh, and N. R. Jennings. Trust in multi-agent systems. *The Knowledge Engineering Review*, 19:1:1–25, 2004.

[33] S. Russel and P. Norvig. *Artificial Intelligence: a modern approach*. Prentice Hall, New York, 1995.

[34] E. A. Schegloff. Reflections on quantification in the study of conversation. In D. P and H. J, editors, *On talk and its institutional occasions*, pages 101–134. Cambridge University Press, Cambridge, 1992.

[35] J. R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3:417–424, 1980.

[36] L. Steels. *The Talking Heads Experiment. Volume 1. Words and Meanings*. Laboratorium, Antwerpen, 1999.

[37] F. Varela, E. Thompson, and E. Rosch. *The Embodied Mind: Cognitive science and human experience*. MIT Press, 1991.

[38] E. von Glasersfeld. Pourquoi le constructivisme doit-il être radical ? *Revue des Sciences de l'Éducation*, 20:21–27, 1994.

[39] E. Wenger. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press, 1999.

[40] M. Wollstonecraft. *A Vindication of the Rights of Woman (Norton Critical Editions)*. W. W. Norton, 1987. Originally published in 1792.

[41] T. Ziemke. Rethinking grounding. In R. . Peschl, editor, *Does Representation Need Reality? Proceedings of the International Conference 'New Trends in Cognitive Science' (NTCS'97)*, pages 87–94. Austrian Society for Cognitive Science, 1997.

[42] T. Ziemke. Are robots embodied? In *Lund University Cognitive Studies*, pages 75–83, 2001.

[43] T. Ziemke. The construction of 'reality' in the robot: Constructivist perspectives on situated artificial intelligence and adaptive robotics. *Foundations of Science*, 6:163–233, 2001.