Emotions in Relation. Epistemological and Ethical Scaffolding for Mixed Human-Robot Social Ecologies

Luisa Damiano †
ldamiano@unime.it

Paul Dumouchel ‡
dumouchp@ce.ritsumei.ac.jp

ABSTRACT

In this article we tackle the core question of machine emotion research – "Can machines have emotions?" – in the context of "social robots", a new class of machines designed to function as "social partners" for humans. Our aim, however, is not to provide an answer to the question "Can robots have emotions?" Rather we argue that the "robotics of emotion" moves us to reformulate it into a different one – "Can robots affectively coordinate with humans?" Developing a series of arguments relevant to theory of emotion, philosophy of AI, and the epistemology of synthetic models, we argue that the answer to this different question is positive, and that it lays grounds for an innovative ethical approach to emotional robots. This ethical project, which we introduced elsewhere as "synthetic ethics", rejects the diffused ethical condemnation of emotional robots as "cheating" technology. Synthetic ethics focuses not on an ideological refusal, but on the concrete sustainability of the emerging mixed human-robot social ecologies. On this basis, in contrast to a purely negative ethical approach to social robotics it promotes an analytical case by case ethical inquiry into the type of human flourishing that can result from human-robot affective coordination.

1. Introduction

"Can machines have emotions?" has generally been taken as the central question in machine emotion research. Can machines feel pain? Can they experience anything? Can they be sad, angry, happy, shameful, exuberant, or

† RG-ESA (Research Group on the Epistemology of the Sciences of the Artificial), Department of Ancient and Modern Civilizations, University of Messina, Messina, Italy.
‡ Graduate School of Core Ethics and Frontier Sciences, Ritsumeikan University, Kyoto, Japan

depressed? Of course, with enough ingenuity and resources, we can make some of them act (or react) as if they had any or all of these emotions and inner feelings; but do they? Do or can machines *have* emotions and feelings? If they cannot, then talk about "machine emotion" is either metaphorical or has to do with the emotions of human beings in their multifarious relations to different artificial systems, especially those that aim at eliciting affective reactions from their users, such as certain robots, virtual agents or computer interfaces. If artificial systems do or can *have* emotions, then the question becomes whether and to what extent these "artificial" or "synthetic emotions" are similar or different from natural (human and animal) emotions.

However, the question "Can machines have emotions?" is not neutral. Underlying this question is a particular understanding of affect which conceptualizes emotion as something which an agent, natural or artificial, *has*. Whether it is conceived as a long term disposition – "this is an angry person" – or as an event – "he was so relieved when he heard she was safe" – emotion is circumscribed as something that happens to the agent, and is hers or his. Emotions in this view are properties in the two senses of the word. First, they are a property that some individuals, i.e. humans, *have* and that other individuals, i.e. artificial agents, do not *have*. Second, emotion is a property of the individual who *has* it: my emotion is mine; it belongs to me. Emotions so understood are fundamentally internal and private. Seen either as "cognitive judgments" (Nussbaum, 2004) or as "affect programs" (Delancey, 2002), emotions take place in an individual mind or body. Even when an emotion coincides with a visible expression, as in blushing, it is the agent's experience, or inner stirring, that is considered to be *what the emotion is*.

Can machines *have* emotions? In the sense defined above, implicit in the way the question is formulated, probably not. However, it is far from clear that this solipsistic understanding of emotion can provide meaningful insights into affective relations, considered either as relations among humans or relations that they establish with artificial agents. Using social robotics as our privileged field of reference, and paradigmatic example, we argue that the question of machine emotion should be addressed in a completely different manner. Specialists in social robotics consider that creating robots able to interact emotionally with humans is central to achieving their field's main goal: building artificial "social partners" for humans (Fong et al., 2003). Thus, rather than implementing emotions as private, internal events, they treat emotions as elements of inter-individual coordination. The question that social robotics

implicitly asks is not "Can robots have emotions?" but "Can robots establish affective coordination with humans?"

This approach expresses a different view of emotions; one which opens a new perspective on machine emotion. This view presents two closely linked advantages with respect to the question: "Can robots have emotions?" First, whether meaningful affective coordination takes place between a robot and its human partner is an empirical or experimental, rather than a purely theoretical or metaphysical issue[1] which, as such, informs and underlies the whole field of social robotics. Interestingly, this new question cannot receive a general "yes/no" answer, as is requested by "Can machines have emotions?", but needs to be addressed over again in relation to different machines and circumstances. The second, closely related advantage concerns ethical reflection on human-robot interactions. A negative answer to the question "Can machine have emotions?", as is commonly the case, leads to the rejection of, or at least to suspicion towards, all technologies that foster affective relations between humans and artificial systems, because, it is argued, they rest on deception. These machines do not have the emotions they pretend to express. This universal condemnation views all such technologies as a danger or evil that needs to be extirpated, or contained through strict rules (Danaher, 2019). As we argued elsewhere (Damiano & Dumouchel, 2018; Damiano, 2020), this attitude tends to isolate ethical reflection and to be self-defeating, for it limits the ethical inquiry on social robots to the definition of general prohibitions or restrictions, which are generally based on pre-existing ethical frameworks that are not specifically focused on this new class of machines. To the opposite, focusing on affective coordination encourages ethics to dialogue with specialists in social robotics, to explore the interactive dynamics that humans develop with the machines they build, and to formulate new and multiple criteria, appropriate for different types of social robots and situations, that is, to the specific characteristics of the different relations humans entertain with these artifacts.

The reminder of this article is dedicated to support this perspective. In section I we show how social robotics, given its methods and objectives, constitutes a frontier of machine emotions. In section II we distinguish and analyze the two main approaches in social robotics, which we view as proposing anew, within this domain of research, the old distinction between strong and

[1]More precisely, the question does not need to be resolved theoretically before it is addressed empirically. In that sense it is experimental, rather than theoretical.

weak AI originally defined by John Searle (1980). In section III we consider social robotics' implementation of emotions, and we characterize it in terms of salient expressions in processes of affective coordination. Section IV draws out the consequence of this alternative view for our understanding of "machine emotion" and the affective competence of social robots. Finally, section V addresses the issue of the ethics of affective relations with robotic social agents.

2. Social robotics: a frontier of machine emotion

Social robots are often viewed in the light of fictional characters such as *The Golem of Prague* (Cohen-Janca, 2017), the Creature from Mary Shelly's *Frankenstein*, or Radius, and the other biochemical robots, staged in *RUR (Rossum's Universal Robots)* by Karel Čapek (1920). However, the forerunners of today's social robots are better represented by a different type of fictional artificial agents, which use their thinking abilities and feelings to be accepted by humans as sympathetic interlocutors, rather than to rebel or even to attempt to exterminate us. One of the earliest of these positive characters is Lester del Rey's *Helen O'Loy* (del Rey, 1938), a fictional robot evincing many features of the machines that contemporary social robotics plans to integrate into human society. In del Rey's short story, two friends modify a housekeeping service robot of the latest model, giving it affective and relational skills. These abilities constitute a central element of the plot, for they dramatically modify the status of the robot. In response to these skills, its human users begin treating Helen O'Loy (for such is the robot's name) as a person, rather than a machine that cooks and cleans. In the eyes of its users, this machine's social and, more specifically, affective competences metamorphose it from an object into a subject, that is, "someone" with "who" a personal, even an intimate relation becomes an option. It is precisely in this sense that Helen O'Loy can be considered as a paradigmatic expression of the goal pursued by current social robotics: building machines that cross over from the purely material realm to our social world. This way, the short science-fiction story about Helen O'Loy illustrates the challenge posed by present-day social robots. This does not lie in the issue of the robots' rights, nor in the possible rebellion of these artifacts – which, for the moment, are very limited machines. As del Rey's short story intelligently anticipates, the challenge of social robots lies in understanding the possibilities and limits of our social interactions with these artifacts, and to

determine and regulate, in a sustainable way, their place within our social ecology – which their presence will inevitably transform.

Only recently have social robots started their voyage from science fiction to engineering. This migration can be traced back to two main events. First, the 1990s "embodiment turn" in cognitive science. The idea that the human mind is an "embodied mind" shifted the focus from disembodied computer programs to complete agents immersed in an environment that is essentially social. From there emerged a conception of intelligent artifacts centered on robots whose cognitive competences are inherently social (Damiano & Dumouchel, 2018). The second crucial event is the 2000s birth of Human-Robot Interaction (HRI), a research area in which the growing interest for the social dimension of artificial embodied intelligence found ideal conditions for development. With the diffusion of advanced robotic appliances operated by humans, HRI emerged as an interdisciplinary domain dedicated to exploring and improving cooperative performances of humans and robots. As a consequence, increased attention was given to the social aspects of human-robot relations, leading to the programmatic notion of robots able to communicate with humans through shared social signals. The project of creating robotic "co-workers" for humans can be seen as the starting point of social robotics. Related research programs, combining embodied AI and HRI, did not simply enlarged the range of potential uses of robots, to include "office, medicine, hotel use, cooking, marketing, entertainment, hobbies, recreation, nursing care, therapy and rehabilitation (...), personal assistance" (Daily et al., 2017). They also enriched the concept of interactive robots, extending it from the notion of human-operated machines to that of social partners. The project became that of enriching robots with "peer-to-peer interaction skills" (Dautenhahn, 2007), enabling them not only to accomplish different tasks in cooperation with humans, but also to interact socially as "peers" (Dautenhahn et al., 2005) – companions, friends, partners.

The strategy adopted by social robotics to create "socially interactive robots" (Fong at al., 2003) can be described in terms compatible with the plot of del Rey's short story: "building machines that are liberated" from "the subject-object dichotomy" (Jones, 2017). In the more technical language of social robotics, this is equivalent to creating robots that have a believable "social presence", defined as the capability of a robotic agent to give its human users the "sense of being with another" (Biocca et al., 2003), or the "feeling of being in the company of another" (Heerink et al., 2008). Doing this entails transferring and adapting to human-robot interactions some aspects of face-to-

face social interaction among humans. Of these aspects, one of the most important is the ability of communicating through emotions.

> The novelty of a mechanical tool that moves with purpose keeps people's attention, but adding emotive capability allows the robot to interact with humans socially. (Daily et al., 2017)

The above quote highlights the central relevance of "machine emotion" for social robotics. Within this domain, research on "affective computing" (Picard, 1997; Picard & Klein, 2002) covers all the areas that are relevant to facilitate appropriate human-robot emotional communication – that is, not only emotion expression, but also emotion sensing, emotion recognition, and emotion generation. Related design and implementation of artificial emotion in robots involve interdisciplinary studies, combining theoretical and experimental HRI with philosophical, psychological, ethological, biological, anthropological, socio-cultural research, as well as design and engineering (Damiano et al., 2015 a). In many cases, the goal is not simply applying knowledge to build believable emotional robotic agents. Social robotics also engages in proper scientific research. Specialists adopt the so-called "synthetic method" (Cordeschi, 2002) – that is, the "understanding by building" research approach (Pfeifer & Scheier, 1999) – to incorporate in robots theories on emotions with the goal of experimentally testing them in HRI settings, and this way contribute to the scientific understanding of emotional processes (e.g., Cañamero, 2005; Damiano & Cañamero, 2010; Asada, 2015). Social robotics is thus a domain where frontier inquiries on machine emotion take place, concerning which the question – "Can machines have emotions?" – is most often raised.

Compellingly, when asked in relation to social robotics this question cannot be considered as merely speculative. What social robotics targets, with the construction of a new class of artificial agents, is the creation of a new type of social relation(s) – namely, human-robot social relation(s). In this case, more than in that of any other engineering domain, the answer to the question of machine emotion is not only philosophical, but also fundamentally practical. The way this question is answered will have a significant impact on the way we will interact and co-exist with social robots. And, if the diffusion of these "social machines" follow the lines of current projections (Brooks, 2002; Šabanović, 2010), our answers to this question will impact our future social and moral evolution.

3. Social robots towards a relational turn

Since its beginning, social robotics has been marked by a tension between two approaches. The first, which arose in the late 1990s, is based on a biologically inspired embodied AI framework. It aims at developing robots that possess social competences in a "substantial sense" (Meister, 2014). This approach targets the production of "socially intelligent" robots whose social presence is grounded in cognition skills modeled on human social abilities. The central idea is that "deep models" of human social competences will allow robots to interpret and to answer appropriately social signals, and thus "show aspects of human style social intelligence" (Fong et al., 2003). The second approach abandons this ambitious goal in favor of a "functional" objective (Jones, 2017), which is to build robotic agents that give only the impression of being socially intelligent. Such robots can be defined as "socially evocative" or "social interfaces" (Breazeal, 2003), and are designed to *simulate* some social competences well enough to engage humans in social interaction. While the first approach wants to create robots that have social skills, the second seeks to make robotic agents whose physical appearance and behavior trigger anthropomorphic projections. In other words, one wants *to endow robots* with social abilities, the other aims *to bring human users* to treat robots *as if* they were genuine social inter-actors.

This tension, opposing the "substantial" and "functional" approaches, is not a novelty introduced by social robotics. It can be considered to reproduce, with a specific focus on robots and social cognition, the opposition between "weak" and "strong" AI that Searle (1980) already recognized within classic (pre-embodiment) artificial intelligence. According to Searle, "strong AI" seeks to create artificial systems that are able to perceive, understand, think, and believe; its goal is to *reproduce* all human cognitive processes. "Weak AI", to the opposite, is based on the idea that artificial systems, at best, can only *simulate* human intelligence; that is, imitate some human cognitive processes. Within social robotics, this distinction is reformulated in terms of the opposition between authentic social skills and mere simulation of human social cognition. The general idea is that, while genuine social abilities are implemented in the internal cognitive architecture of robots, their simulation rests on external anthropomorphic expressions of robotic agents. Just as strong AI seeks to create artificial minds, in social robotics the substantial approach envisions the generation of an artificial social species, whose members will integrate human society (e.g., MacDorman & Cowley, 2006). Just as weak AI sees in computer

programs not artificial minds, but useful artifacts, functional social robotics considers its robotic agents not as authentic social partners, but as acceptable interactive technological solutions (e.g., Duffy & Zawieska, 2014).

Interestingly, this distinction recently reappeared in a new guise within "android science", a sub-division of social robotics and HRI that is dedicated to the production of increasingly human-like robots (MacDorman & Ishiguro, 2006; Ishiguro, 2016). Here, Searle's strong/weak AI opposition is used to develop a framework to assess the "human-likeness" of current and future humanoid robots (Kahn et al., 2007). More specifically, this evaluative approach adopts the strong/weak duality to qualify the "force" of "ontological" and "psychological" assessments of robots' human-likeness. The first form of evaluation concerns the ontological status of humanoid robots – "are humanoid robots humans or machines?" The second focuses on the perception of their status – "are humanoid robots *perceived* as humans or as machines?" Strong ontological claims assert that humanoid robots "will become" or "are" humans, something which some proponents of android science consider to be possible in the future. Weak ontological claims deny this to be the case. Analogously, psychological claims are "strong" when humans see human-like robots as other humans, while they are "weak" when humans regard them as mere machines.

According to us, the most interesting aspect of this reformulation is that it brings back into focus – beyond Searle's criticism of the Turing test – the role of human perception in the assessment of artificial systems. Within social robotics, the growing attention towards the users' evaluation can be viewed as a general trend, to the extent that the literature characterizes it as a paradigm shift. Raya Jones (2017), for example, describes this orientation as an increasing focus on users' experience that leads specialists to assess robots' "sociability" primarily – or even exclusively – in terms of the users' evaluation of their social interaction with these artificial agents. Jones interprets this as a "subtle 'paradigm shift'" in the field, which brings specialists to implicitly dismiss the diffused idea that sociability corresponds to a set of individual skills or competences that can be re-enacted, deeply or superficially, by machines. As Jones emphasizes, current procedures assessing robots' sociability do not consider that this ability is an ensemble of traits characterizing the artifacts as individual agents. Evaluations do not focus on the robots' social features, but on the users' perception of their artificial partners, suggesting that robots' artificial sociability arises in, and is a property of, human-robot interaction.

While we agree with Jones that the diffusion of user-centered approaches can be seen as a marker of a "paradigm shift", we doubt that this transformation in itself can be described as a "relational turn" (Jones, 2017). The increased focus on users' experience and judgment indicates that specialists in social robotics tend to locate robots' sociability more in the users' eyes than in the relation between users and robots. This perspective, together with the related *modus operandi*, can be considered only as a first step towards a relational turn, since – and only to the extent that – it involves the users in the generation of robots' sociability. Yet, we argue, a proper relational turn requires another step, which changes radically the angle on such a sociability. It demands that the social character of robots, rather than reduced to a users' property (and projection), be recognized as a *distributed* property. That is, one which emerges from the interactive dynamic taking place between users and robots. A property that can neither be implemented as a trait of individual robots, nor merely understood as their users' projection, because it is distributed in the mixed human-robot system that users and robotic agents together form through their interactions.

While we take a different perspective of the meaning of this "relational turn", we agree with Jones that there are indications that a paradigm shift is actually taking place, at least in frontier sub-divisions of social robotics focusing on machine emotion.

4. Robotic emotion as relation

From the late 1990s, the tension between substantial and functional approaches to social robotics appeared in a specific version, more narrowly focused on emotions, which can be defined as the contrast between the "internal" and "external" approaches to the "robotics of emotion" (Damiano et al., 2015b; Dumouchel & Damiano, 2017; Damiano & Dumouchel, 2018). The first orientation, "substantial" or "internal", focuses on endowing robots with deep models of human or animal emotional systems. The second, "functional" or "external", produces robots that express emotions, but do not have any internal mechanism or architecture designed to generate emotional processes. As such, this division can be considered to reproduce, within the robotics of emotion, Searle's classic distinction between strong and weak AI. Indeed, whereas internal robotics of emotion pursues the creation of "genuine emotion", external robotics targets the mere "simulation of emotion".

However, the structure of the field is not reducible to Searle's dichotomy. Interestingly, the robotics of emotion presents a third, growing research trend, which tries to combine aspects of the internal and of the external approaches in order to overcome their limitations. Mainly these are, respectively, the extreme difficulty of implementing "deep models" of emotion in robots, and the lack of credibility of purely expressive robots. Driven by the goal of developing real-time, long-lasting affective interactions between human and robotic agents, this new trend, often called the "affective loop approach", focuses neither on the robots' ability to generate emotions, nor on their emotional expressivity. Rather, as its name suggests, it concentrates on the robots' capacity to engage humans into affective exchanges. Its objective is to bring "the user to [affectively] respond [to the robotic system] and step-by-step feel more and more involved with the system" (Höök, 2009), and in such a way to favor continued human-robot social interaction. To achieve this, specialists attempt to endow robots with "intelligent expression" (Paiva et al., 2014), that is, emotional expression dynamically coordinated with that of their users. This is attained by articulating realistic expressive systems with mechanisms that regulate their functioning in order to meaningfully couple the robotic and the human agents' emotional manifestations. In the most advanced models, this is implemented through internal architectures that include not only emotion perception and recognition, but also emotion generation mechanisms, which are typically only loosely modeled on animal or human emotional systems, or even "invented" for the robots.

We believe that this approach to machine emotion is effectively realizing the *relational turn* in social robotics. While both internal and external robotics of emotion focus on robots as individual agents, the affective loop approach is centered on human-robot pairs or groups. Whereas internal and external robotics of emotion aims at constructing (genuine or simulated) emotions in single agents, the goal sought by the affective loop approach is for emotions to arise within the human–robot system as a whole. What it targets is not an individual emotional process, confined within the robotic or the human agent, but a distributed emotional dynamic, grounded in the recursive coordination of the affective expression of the robotic and human agents involved in the affective interaction.

This approach suggests an implicit – for in the field it remains unformulated – conception of emotions completely different from viewing them as individual properties or skills. Social interactions, whatever else they may be, clearly are

joint ventures. Their success or failure, their permanence or fugacity, depend on all the partners involved, and that characteristics of social interactions is central to social robotics. The emotions that emerge in social relations and constitute a fundamental and integral part of the interaction, should in the same way be viewed as collective creations, rather than as private, individual events. This idea, that emotions are joint creations emerging in affective coordination, challenges our common understanding of emotions in a radical way.

Affective coordination, we argue, is the process through which we jointly determine our reciprocal intentions of action and intentions towards each other (Dumouchel, 1999, 2008). Game theoretical models of coordination in economics or biology analyze how two or more agents, whose interests at least partially converge, coordinate their action in view of a particular goal that they share. Because they assume that the convergence of interest is sufficient to insure that the agents will collaborate, these models never ask the questions: "who do you want to play with?" and "who don't you want to play with?" Yet those questions are fundamental in both human life and social robotics. In social robotics, because a successful artificial social agent is one "we want to play with" independently of whatever task we may engage in together. In life, because the repulsion or pleasure that we find in interacting with others brings us to weigh quite differently the interests whose convergence or divergence the theory of games takes as given.

Affect coordinates persons with each other. The issue is not that of coordinating the actions of multiple agents so that they can successfully achieve a common goal, but to determine whether or not they want to pursue a common goal. Affective interaction is the way through which we coordinate our reciprocal intentions of action. It is thus prior to both cooperation and conflict, or competition, as well as to lofty or lazy indifference. For members of a species like ours, whose advantages and disadvantages primarily depend on interactions with each other, rather than on solitary relations to the natural environment, material situations under-determine agents' interests. Individuals' preferences for this or that outcome are conditional upon other agents' preferences and vice versa. Affective coordination is the mechanism through which these preferences are determined and is lifted the uncertainty concerning the other's intention towards me and my intention towards her or him.

It is a reciprocal process in which a first agent partially determines the intentions of a second towards her or him, while this second partially determines the first's intention towards the second. To put it otherwise, your attitude

towards me partially determines my attitude towards you and, vice versa, my attitude towards you partially determines your attitude towards me. In this reciprocal process, affective expressions occupy the central place. The process is, however, generally unconscious. We are mostly unaware of the how our affective expression affects other, or even of what – gestures, postures, tone of voice – constitutes this expression. If we attend more to the affective expression of others, it nonetheless generally remains below the level of conscious apprehension. The process of affective coordination is thus unattended and non-intentional in both the ordinary and the technical sense of that term. It is neither a conscious object, nor something that I want to do. Though my intention towards you results from this process, and vice versa, it cannot itself be intentional under pain of infinite regress.

Emotions, we propose, are moments of this continuous process of coordination which become salient for one or more of a variety of reasons. For example, because they are fixed points of coordination, because they represent a sudden reversal, because of lasting inability to reduce uncertainty concerning the other′s intention, etc. In every case, what determines a moment to be an emotion is not an individual′s mental or bodily state, but some aspect of a process of coordination that involves at least two agents.

Such a proposal entails a profound transformation of the concept of emotion. Rather than being defined as something that happens to an individual, either an internal event characterized by some mental or bodily change, or a public manifestation that is shaped by, or even created as a result of, social rules and expectations, an emotion is here construed as a relational property. Just like being "taller than" is a property of an individual agent, but which that agent, for example John, cannot have by himself, independently of his relation to others, say to Louise or Peter. Similarly, "being sad" is certainly a property of an individual agent but, if it is a salient moment in a process of affective coordination, then it is not one she or he can have independently of others. *It does not follow from this that another person is necessarily the cause or "target" of the emotion*, for that does not need to be the case in order for a moment to be salient in a process of affective coordination.

One of the evident consequences of this re-conceptualization is that emotions so defined do not perfectly coincide anymore with our everyday use of the term. Some events which ordinary language identifies as emotions are not according to this definition and alternatively others, which it fails to recognize as such, may very well be salient moments in the process of affective

coordination. This, as far as we are concerned, is all for the best. A central weakness of current theories of emotions, and the main reason why they have failed so far to propose agreed upon characterizations of emotions, or even a common list of basic emotions, is that they tend to consider that the everyday term "emotion" properly identifies and provides basic incontrovertible insight into the phenomena which they want to investigate. There is however little reason to believe that to be the case. Little reason to think that our spontaneous grasp of emotions is closer to a scientific understanding of the phenomena they point to, than our spontaneous understanding of physics is to the scientific discipline that bears that name (Wolpert, 1992). Our proposal goes against common sense; that is its force, we argue, not an objection. It is further perfectly well adapted to the challenge that social robotics faces: understanding and bringing about emotions in relation.

Affective coordination rests on affective expression and given that emotions are salient moments in this process, it follows that expression precedes emotions. An emotion then should not be seen as a cause of which affective expression is an effect, but rather as an effect of a dynamic of reciprocal affective expression. Because an emotion is not essentially an internal event, the question is not whether an agent, natural or artificial, *has* the relevant internal event relative to his (or its) expression – "do machines *have* emotions?" – but whether this moment corresponds to a *projectible* predicate (Goodman, 1954) of the relation. As one of us argued elsewhere, "sincerity is consistency", a fixed point in a dynamic of relation (Ross & Dumouchel, 2004).

5. Social robots: models or partners?

The affective coordination hypothesis, applied to our exchanges with social robots, argues that human-robot emotional interactions, in spite of their irreducible differences with human-human (and human-animal) interactions, also include forms of affective coordination. As described in (Dumouchel & Damiano, 2017; Damiano & Dumouchel, 2018), the limited and rigid character of current human-robot emotional exchanges does not exclude the possibility of genuine affective relationship. According to us, narrowness and rigidity define only one aspect of human-robot affective relations. Social robots can nonetheless function as affective agents, rather than simply as "emotional regulators", like stuffed toys or comfort food.

This position is controversial. A most common objection raised against it is that robots, like puppets, do not have emotional states. Lacking an internal animal-like emotional system, these artifacts are unable to reciprocate emotions. It is argued that, in consequence, we cannot establish bi-directional affective relations with them. Whether or not our affective exchanges are with social robots that are designed following the affective loop approach is beside the point, is it claimed. Robots are inert objects, incapable of emotions. Like dolls, they are not true emotional agents, though we may become attached to them.

We do not deny that robots are objects, nor that social robotics uses insights from puppetry to design them as technologically enhanced dolls to induce in users a "suspension of disbelief", making them appear as humans' peers (e.g., Duffy, 2006; Duffy and Zawieksa, 2012; Zawieksa & Duffy, 2014). Our claim concerns neither their ontological status – "are they objects or subjects?" – nor the users' perception of robots – again as either objects or subjects. The affective coordination hypothesis focuses on the interactive dynamics that arise between humans and social robots, and the interactive roles that these machines can play. We propound that, during human-robot interactions, these "sophisticated dolls" can reproduce key elements of the dynamics of human-human (or human-animal) affective coordination. More in detail, they can reproduce, together with their human users, aspects of the recursive coordination of affective expressions that leads inter-actors to mutually determine their disposition to action. This is the sense in which, in our view, these artificial agents become affective inter-actors, or emotional agents: a new kind of partners in affective relations. Compared to humans or other animals, they have significant limits as emotional agents, as do the affective relations which we can establish with them. However, robot-human affective interactions are nonetheless authentic because robots can participate in affective coordination dynamics with humans.

Current epistemological reflection on "artificial" or "synthetic models" may help us clarify and support this position. Models produced through the "synthetic method" or "understanding by building" approach are not merely abstract theoretical representations. They incorporate theories in biological and/or cognitive processes to allow the experimental exploration of these processes. In that sense, they are *material models* of the target process. They are built either as robotic, computational, chemical, or mixed functioning systems, and are used to test the ability of our scientific theories to generate the target processes, that is to say, to explain operationally their genesis in natural

scenarios (Damiano et al., 2011). More generally, such models are used to investigate experimentally aspects of the target processes that are not accessible in traditional research scenarios. Following this approach, for example, specialists in synthetic biology build synthetic models of cells – "synthetic cells" – for a wide range of research purposes, including: testing the capability of our current theories at recreating the basic processes through which the first biological cells were constituted; reproducing and studying some of the complex activities of biological cells; experimentally manipulating and exploring natural communicative processes among bacteria, such as *quorum sensing* (Damiano & Stano, 2018; Rampioni et al., 2018).[2]

One of the main issues in the epistemology of synthetic models is their relevance for the scientific understanding of the target process (e.g., Webb, 2001). A recent proposal (Damiano & Cañamero, 2010; Damiano et al., 2011) introduces two "criteria of relevance" for synthetic models. These can help clarify why we think that social robots participate in affective coordination dynamics when interacting with humans.

The first criterion aims at assessing the "phenomenological relevance" of synthetic models. A synthetic model is phenomenologically relevant if it exhibits, within strict parameters, the same phenomenology as displayed by the target system, regardless of the particular mechanisms used to achieve this. The focus is on the capacity of an artificial system to reproduce the behavior of a natural system, independently of how that behavior is generated. The second criterion evaluates the "organizational relevance" of synthetic models. A synthetic model is organizationally relevant if its organization reproduces the target system's organization, according to a pertinent scientific theory. This second criterion shifts the focus from phenomenology to organization, that is, from the natural system's behavior to its underlying mechanisms. On this basis, in the context of robotic modeling of emotions, a model is "phenomenologically relevant" if it reproduces the affective behavior observable in the target systems. It is "organizationally relevant" if it reproduces the mechanisms that underlie that behavior.

One advantage of these criteria is that they allow us to escape the rigid alternative between pure imitation and genuine reproduction. Indeed, they

[2] In a way, synthetic systems of this kind, like synthetic cells, are models in the sense Suppe's (1989:167) *iconic model*: "an entity that is structurally similar to the entities in some class". On this basis, they are also models in the sense of "a new car model", as they instantiate a particular type or kind.

enable a classification of synthetic models (Damiano et al., 2011) that takes into account various combinations and different forms of phenomenological and organizational relevance. From this epistemological point of view, the "weak/strong AI" dichotomy is a simplifying approach, which corresponds only to the two extreme, or "pure", cases. Weak AI propounds a purely imitative, phenomenologically relevant models of intelligence. While strong AI aims at models that are perfectly organizationally relevant, *and* that, as a consequence, are phenomenologically relevant also. The underlying hypothesis here is that organizational relevance necessarily brings about phenomenological relevance, implying a hierarchy between the two poles of this classic dichotomy. To the opposite, including all the combinations of organizational and phenomenological relevance of models involve other relations between them than a simple hierarchy. For example, there is the possibility of synthetic models which are organizationally relevant *only*, in the sense that the reproduction of the target's organizational mechanisms gives rise to new behaviors, different from those of the target system. In this sense, the classification proposed in (Damiano et al., 2011) leaves open the possibility that synthetic modeling may lead to *man-made variants* of natural systems that manifest new behaviors. This classification also allows to distinguish between "progressive" and "interactive" forms of phenomenological or organizational relevance. "Progressive relevance" refers to artificial systems that exhibit unexpected behaviors which are pertinent to the inquiry on the target processes. "Interactive relevance" characterizes artificial systems able to engage natural systems in dynamics germane to the scientific investigation on the target processes conducted through these models[3].

Using this classification to assess synthetic models is not a straightforward operation. For example, assessing robotic models of emotion gives different results depending on the notion of emotion adopted as reference. Among other things, because the target phenomena and mechanisms will vary when using different concepts of emotion. For example, the individual conception of

[3] This characterization of two criteria of relevance and the related classification of forms of relevance develops work originally introduced in (Damiano & Cañamero, 2010; Damiano et al., 2011). These developments were presented at the workshop SA-BCS 2018 at ALIFE18 (SA-BCS 2018, L. Damiano, *The Synthetic Method. Proposing a framework to the synthetic sciences of life and cognition*).

emotion and the affective coordination approach focus on different behaviors and different underlying organizations.

On the classic individualist view of emotion robots are models of "individual affective agents". That is, agents whose emotional processes are generated by private internal mechanisms that are proper to each agent. From this point of view, all existing social robots have a very low, or even no organizational relevance at all. Specialists of the internal robotics of emotions, who generally adopt such a view of emotions, unsurprisingly consider that reaching significant organizational relevance is "a task for the future" (Parisi, 2014). Most current social robots are artifacts that, like Paro or Keepon, do not reproduce animal-like or human-like emotional mechanisms or physiology, but merely display some limited form of emotional behavior. Interestingly, all of these robots nonetheless demonstrate *interactive phenomenological relevance*. As many studies show (e.g., Kahn et al., 2002; Severson & Carlson, 2010; Gaudiello et al., 2015), through their ability to reproduce important aspects of animal or human emotional behavior, these robots succeed, to different extents, in engaging humans in affective interactions. This interactive phenomenological relevance, viewed through the eyes of classic theories of emotion, defines social robots as purely *imitative models.* They allow us to explore emotional processes in natural systems; that is, they can be used experimentally to study and to manipulate interactive emotional dynamics in humans and animals. As we will see later on, manipulation is precisely what is commonly reproached to them from an ethical point of view once they leave the lab and enter social life.

When assessment is based on the affective coordination thesis, the focus is on the robots' ability to model inter-individual emotional dynamics, rather than individual emotions. The target phenomenology is different. It is not individual internal events, but interactive behaviors. Similarly, organizational relevance is not to be measured against the internal mechanisms that are deemed to produce intra-individual emotional processes. Rather, organizational relevance is now to be assessed in relation to the central mechanism of affective coordination: the recursive process of co-expression that leads inter-actors to co-determine their dispositions to action. Considered from this angle, many of the current social robots appear able to successfully reproduce at least some key aspects of this mechanism. In fact, many of these robots can engage humans in recursive dynamics of emotional expression that trigger in their users – and sometimes

also in the robots themselves[4] – changes in their disposition to action. From this point of view, these social robots appear as synthetic models of agents in affective coordination; models to which we can ascribe partial organizational *and* partial phenomenological relevance. From an ethical point of view also things now look different, for what is involved is not manipulating the emotions of individual agents, but participating to an inter-individual, shared affective dynamic.

Currently, the emotional expression of robots, the range of actions of which they are capable, and thus the complexity of the forms of interaction and coordination available to them remain quite limited. Therefore, at this point, affective coordination with them has neither the wealth nor the breath of what we find in human-human, or even human-animal, affective interactions. [5] However, within these limits, some social robots function successfully as models of human emotional inter-actors, although the internal mechanisms that support their capacity to participate in affective coordination dynamic are quite different from those of their human partners.

At this point, there are two epistemological options. We can either focus on the *partial*, or limited, *organizational and phenomenological relevance* of these robotic models of affective coordination, viewing social robots as *models of "natural" – human or animal – affective agents*. Alternatively, we can focus on the *synergy* of partial organizational *and* partial phenomenological relevance revealed by social robots. In this case, their difference from "natural affective agents" defines them as a *new* type of affective agents, and the phenomenology of human-robot affective coordination as a *new and different* emotional phenomenology.

Should we consider social robots as models of natural affective agents in coordination, or as new affective partners for humans? The interest of this question is not merely speculative. The importance of distinguishing and choosing between these two options is that they lead to different ethical stances, which have significant impacts on the future of our social ecology.

The first option remains close to the classic individualist view of emotions and encourages a popular ethical stance. According to this perspective, once social robots move out of the labs, they can no longer be defined as models of

[4] This depends on the type of machine we are dealing with.

[5] However the range of affective coordination of some semi-autonomous robots, for example KASPAR, can be quite large. See (Dumouchel & Damiano, 2017: 158-163).

emotional processes. Within our social ecology these robots constitute a form of "cheating technology", which induces in humans the illusion of reciprocal affective relations. This attitude, currently dominant in ethical reflection (Danhaer, 2019; Damiano & Dumouchel, 2018; Damiano, 2020), leads to rejecting all social robots, because the fault lies in the design of the technology itself. Based on this perspective, some authors even suggest to proscribe all social robots, independently of their characteristics and concrete uses (e.g., Turkle, 2010). A radically different ethical position emerges from the affective coordination approach, once we accept robots as a new type of affective partners. The core of this different approach lies in recognizing that emotional interactions with social robots constitute a specific form of affective coordination. From this point of view, just as it is fundamental not to confuse artificial intelligence with human intelligence – and not to see the first as superior, "more of the same" – it is important not to confuse human-robot affective coordination with human-human affective coordination – seeing the first as inferior, "less of the same" (Dumouchel, 2019). Rather we need to recognize that we are dealing with a different but related phenomenon that requires a new approach.

6. Synthetic ethics

As mentioned earlier, the ethics of affective relations with social robots is dominated by the claim that these can never be authentic, because machines do not have emotions. Whatever sentiments humans may harbor towards a machine, they will never be reciprocated. Such a relation, in the end, will inevitably reveal itself to be a fraud. This approach, paradigmatically expressed by Sherry Turkle (e.g., 2010; 2011), is often associated with a dystopian view of the diffusion of social robots. Among their expected negative effects we find: the delegation of care and support relationships to machines, resulting in the exclusion from the social sphere of subjects with special needs; the manipulation of vulnerable individuals through the illusion of reciprocal affective relations with robots; the development of a preference for social and emotional interactions with robots rather than humans, leading to unprecedented levels of social isolation; gaps in the cognitive and social development of new generations, and so on. Not surprisingly, then, social sustainability, argues Turkle, requires "the exclusion of these evocative objects [social robots] from the realm of our relationships" (Turkle, 2010). This position has the merit of recognizing some of the dangers

inherent to social robots. However, its proposal appears more harmful than helpful (Damiano & Dumouchel, 2018; Damiano, 2020).

One evident unhelpful implication of this general condemnation is that it excludes *a priori* any possible beneficial uses of social robots. For example, robot assisted therapies that improve the social skills of children with special needs (e.g., Cabibihan et al., 2013; Lehmann et al., 2014), or socially assistive robotics projects, dedicated to supporting vulnerable persons (e.g., Tapus, 2007). However, this position's main weakness is that it insulates ethical reflection from scientific research and technical developments. This generalized rejection of social robots renders ethics unable to offer guidelines to help robotic research reduce the risks and/or maximize the benefits of specific projects. Furthermore, this total condemnation is destined to be unheeded by the scientific community, and likely to have a negative influence on the development of social robots, even from an ethical point of view. Indeed, it can only block, never help, social robotics projects that need ethical approval, which are typically dedicated to building social robots apt to engage with humans in socially meaningful ways, as therapists, trainers, mediators, and caregivers.

A good example of the counter-productive dimension of this negative approach comes from entertainment robotics. In this domain, some projects encourage users to exercise violence against social robots, of which sex robots with an integrated "rape option" constitute an extreme case (Damiano & Dumouchel, 2018).[6] The logic underlying the general condemnation ethics implies that these developments merely involve "simulated violence", since the violence cannot be more real than the "simulated suffering" of the robot against which it is exerted. Thus, with regard to violent entertainment robotics, this ethical approach unwittingly provides an implicit legitimization, that it is at lost to refute. This is not an unfortunate accident. By condemning as "false" or "inauthentic" our emotional and social interactions with robots, this approach, on one side, leads users to view their own actions as unreal and without consequences, and, on the other, fails to engage in a critical exploration of the influence that interaction with social robots can have on human ethical behavior. Its likely outcome, far from being the proscription of social robots, is that social robotics will develop in the absence of focused ethical research, and will produce its own ethical guidelines that in the majority of the cases, will likely merely reflect common prejudices and economic interests.

[6] https://www.nytimes.com/2017/07/17/opinion/sex-robots-consent.html

Clearly there is room for a different ethical approach: One that concretely engages in supporting the sustainable social development of social robots, which is distinct from the ideological rejection of any mixed human-robot social ecology.

There is need for an ethical investigation that actively participates in the processes of ideation, design, and construction, as well as of introduction of social robots in human social contexts. An approach which, instead of considering human-robot emotional and social interactions as resting on "simulation" and "falsehood" on the part of artificial agents, recognizes that both humans and robots participate in a dynamic of affective coordination that can significantly affect their conduct. This perspective has to be at the heart of ethical inquiry if it is to engage in critically studying and assessing – case by case, project by project – the effects of human-robot affective coordination, and to cooperate with practitioners of social robotics to define concrete, specific ethical guidelines that can be implemented. We previously introduced, under the name of "synthetic ethics", this approach that we are presently developing on the basis of the relational view of emotion.

Synthetic ethics does not only abstain from the general condemnation of social robots. Unlike many other ethical approaches to robots (and technological innovations in general) it is centered, neither on measuring the social impact of technological changes – as utilitarian approaches tend to do – nor on determining rules to control the ethical behavior of robots – which both utilitarian and deontological approaches aim to do. Not that we dismiss or undervalue such objectives. However, *synthetic ethics* pursues a different one. It draws part of its inspiration from the Aristotelian idea of prudence, or φρόνησις, which implies openness to the unexpected, and recommends careful attention to what is new.[7] The goal of *synthetic ethics* therefore is not to judge of what is new or of changes according to a finite set of pre-existing rules, but to discover new rules as needed. Rules that are specifically formulated in accordance with novel developments, in particular, the development of human-robots social affective relations.

We named this approach *synthetic ethics* to emphasize its closeness to "learning by doing" and the "synthetic approach" characteristic of the

[7] Cf. Aristotle, 2004. Other perspectives on the role that Aristotelian phronesis can play in ethical research on ICT and social robots are presented in Ess (2007) and Polak & Krzanowski (2017).

methodology of social robotics. The reference to the "synthetic method" is, however, not limited to highlighting the intention to cooperate with social robotics, by studying in laboratories social and affective human-robot interactions. The ambition is also to actively support the sustainable development and social diffusion of these robots.

Identifying the dangers they present should not be seen as a cause for rejection, but as research questions that need to receive concrete solutions that can be implemented. How can we build social robots that operate as "social connectors", strengthening the social bond instead of weakening it? How can robots be used to facilitate supportive relationships with people who have special needs? How can we make social robots able to stimulate, rather than damage, the cognitive, social and emotional development of their humans partners? Although this approach is still in its infancy, its proactive inclination has already encouraged concrete synergies between philosophical reflection and scientific-technological research, which promises effective applications (Rajaonah et al., 2020; Ocnarescu & Scimma, 2020; Cappuccio et al., 2019). Furthermore, the reference to the synthetic approach wants to highlight another goal of *synthetic ethics*: the use of experimental scenarios structured by social robotics to study and better understand humans as ethical agents. In other words, using these new artifacts to deepen our self-knowledge from an ethical point of view, with a focus on the choices and ethical conduct that emerge from human-robot interactions. How does interacting with robots transform our ethical behavior? How do choices and values change in the context of these new forms of affective coordination?

This way, s*ynthetic ethics*, as an ethical undertaking based on the synthetic approach and a relational view of emotion, generates a concrete alternative to the ideological rejection of the arising human-robot social ecologies: Grounding their sustainability in a productive circuit between the process of self-knowledge and the process of self-transformation that we can realize through interactions with social robots.

REFERENCES

Aristotle, *The Nicomachean Ethics*. Penguin Classics, 2004.

Asada M. (2015). Towards Artificial Empathy, *International Journal of Social Robotics*, 7, 19-33.

Biocca, F., Harms, C., and Burgoon, J. K. (2003). Toward a more robust theory and measure of social presence. *Presence* 12, 456–480.

Breazeal, C. (2003). Toward sociable robots. *Rob. Auton. Syst.* 42 (3), 167–175

Brooks, R. (2002). *Flesh and Machines: How robots will change us*. Vintage Books, New York

Cabibihan, J.-J., Javed, H., Ang Jr., M. Aljunied, S.M. (2013). Why Robots? *International Journal of Social Robotics*, 5 (4), pp. 593-618.

Cañamero L. (2005). Emotion understanding from the perspective of autonomous robot research, *Neural networks*, 18, 4, 115-148.

Cappuccio, M., Peeters, A., McDonald, M. (2019). Sympathy for Dolores. Philosophy and Technology, https://doi.org/10.1007/s13347-019-0341-y

Cohen-Janca, *The Golem of Prague*, 2017.

Cordeschi, R. (2002). *The Discovery of the Artificial.* Alphen an den Rijn: Kluwer.

Danaher, J. (2019). The Philosophical Case for Robot Friendship. *Journal of Posthuman Studies*, 3 (1), 5-24.

Damiano L. (2020). Mind, robots and mixed social ecologies. Towards an experimental epistemology of social robots, *Sistemi Intelligenti*, XXXII, 1, 27-39.

Damiano L. & Cañamero L. (2010). Constructing Emotions. In *AI Inspired Biology*, SSAISB, 20-28.

Damiano L. & Dumouchel P. (2018). Anthropomorphism in Human-Robot Co-Evolution, *Frontiers in Psychology*, doi: 10.3389/fpsyg.2018.00468.

Damiano L., Dumouchel P., Lehmann H. (2015a). Artificial Empathy: An Interdisciplinary Investigation, in *International Journal of Social Robotics, 7.*

Damiano L., Dumouchel P., Lehmann H. (2015b). Towards Human-Robot Affective Co-Evolution, in *International Journal of Social Robotics*, 7, 7-18.

Damiano L., Hiolle A. and Cañamero L. (2011). Grounding Synthetic Knowledge, in Advances in Artificial Life, ECAL 2011, MIT, 200-207.

Damiano, L. & Stano P. [2018], Synthetic Biology and Artificial Intelligence. In Complex Systems, 27, 3, 199-228.

Daily SB, James MT, Cherry D., Porter III JJ, Darnell SS, Isaac j., Roy T. (2017). Affective computing. In Jeon M. (Ed). Emotions and Affect Fctors in HCI, Academic Press, 213-231.

Dautenhahn K. (2007), Socially intelligent robots: dimensions of human–robot interaction, *Philosophical Transactions B*, 362(1480), 679–704.

Dautenhahn K., Woods S., Kaouri C., Walters M. L., Koay K. L., Werry I., What is a Robot Companion – Friend, Assistant or Butler?

Delancey C. (2002) *Passionate Engines: What Emotions Reveal about the Mind and AI*. Oxford University Press.

del Rey, L. (1938). Helen O'Loy, in *Astounding Science Fiction*, December.

Duffy, B.R. (2006) Fundamental issues in social robotics. *International Review of Information Ethics* 6, 31-36.

Duffy B.R. and Zawieska K (2012) Suspension of Disbelief in Social Robotics. 2012 IEEE RO-MAN, Septemer 9-13, 484–489.

Dumouchel P. (1999). *Émotions*. Paris, Institut Synthélabo.

Dumouchel, P. (2008) « Social Emotions » in Canamero L., Aylett R. (eds), *Animating Expressive Characters for Social Interaction*, John Benjamin Publishing Company, Amsterdam/Philadelphia 2008, 1-19.

Dumouchel P. (2017) Acting together in dis-harmony. Cooperating to conflict and cooperation in conflict. in *Studi di Sociologia* (2017) 4, 303-318

Dumouchel, P. (2019) Intelligence, Artificial and Otherwise *Forum Philosophicum* 24, 2, 241-258.

Dumouchel P. & Damiano L. (2017). *Living with Robots*. Harvard University Press.

Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. *Rob. Auton. Syst.* 42, 143–166.

Gaudiello, I., Lefort, S., and Zibetti, E. (2015). The ontological and functional status of robots. Comput. Human Behav. 50, 259–273.

Goodman N. (1954) *Fact, Fiction and Forecast*, Harvard University Press.

Hawkley L. C. & J. T. Cacioppo (2010). Loneliness Matters: A Theoretical and Empirical Review of Consequences and Mechanisms in *Annals of Behavioral Medicine* 2010. Oct 40(2): 218-227.

Heerink, M., Kröse, B., Evers, V., and Wielinga, B. (2008). The influence of social presence on acceptance of a companion robot by older people. *J. Phys. Agents* 2, 33–40.

Hobbes, Th. (1651/1976) *Leviathan* C.B. Penguin Books .

Höök, K. (2009). Affective loop experiences: designing for interactional embodiment. *Philos. Trans. R. Soc. B*364, 3585–3595.

Ishiguro H. (2016) Android Science. In: Kasaki M., Ishiguro H., Asada M., Osaka M., Fujikado T. (eds) Cognitive Neuroscience Robotics A. Springer.

Jones, R. (2017). What makes a robot 'social'? *Social Studies of Science*, 47 (4), 556-579

Kahn, P. H., Ishiguro, H., Friedman, B., Kanda, T., Freier, N. G., Severson, R. L., Miller, J. (2007). What is a human? *Interaction Studies* 8:3, 363–390.

Kahn, P. H., Friedman, B. Jr., and Hagman, J. (2002). 'I care about him as a pal': conceptions of robotic pets in online AIBO discussion forums, in *Proceedings of the Extended Abstracts at the Conference on Human Factors in Computing Systems*, New York: ACM Press, 632–633.

Lehmann, H., Iacono, I., Dautenhahn, K., Marti, P., Robins, B. (2014). Robots companions for children with Down Syndrome. Interaction Studies , 15 (1), 99-112.

MacDorman and Cowley (2006). Long-term relationships as a benchmark for robot personhood. Proceedings of the 15th IEEE *International Symposium on Robot and Human Interactive Communication*, 378-383.

MacDorman, K. F., and Ishiguro, H. (2006). The uncanny advantage of using androids in social and cognitive science research. *Interaction Studies* 7(3), 297-337.

Meister, M. (2014). When is a Robot really Social?An Outline of the Robot Sociologicus Science, Technology & Innovation Studies, 10, 1, 107-134

Nussbaum, M. (2004) *Upheavals in Thought*. Cambridge University Press

Ocnarescu, I., & Sciamma, D. (2020). Homes through the Design Shift in the Digital Age. In A. Argandona, J. Malala e R. Peatfiled (eds), The Home in the Digital Age, Routledge (in press)

Paiva, A., Leite, I., & Ribeiro, T. (2014). Emotion modeling for social robots, in *Handbook of Affective Computing*; eds R. Calvo, S. D'Mello, J. Gratch, and A. Kappas, Oxford, Oxford University Press

Parisi, D. (2014). *Future Robots*. London, John Benjamins.

Pfeifer, R., & Scheier, C. (1999). *Understanding Intelligence*. Cambridge, MIT Press.

Picard, R.W. (1997). Affective Computing, Cambridge, MIT Press.

Picard, R.W., & Klein, J., 2002. Computers that recognise and respond to user emotion. *Interact. Comput.* 14, 141–169.

Rajaonah, B. Sarraipa, J, Wallard, L., Huftier, A., Abed, M. (2019). Affective Social Robots and Ethics. Manuscript under review.

Rampioni G., Leoni L., Mavelli F., Damiano L., Stano L. [2018], Interfacing Synthetic Cells with Biological Cells: An Application of the Synthetic Method, in Ikegami et al., Alife 2018, MIT Press, 145-146.

Ross, D. & P. Dumouchel (2004) Emotions as Strategic Signals in *Rationality and Society* 16.3, 251-286.

Šabanović, S. (2010). Robots in society, society in robots. *International Journal of Social Robotics*, 2(4), 439–450.

Searle, J. R. (1980). Minds, Brains and Programs, *Behavioral and Brain Sciences*, 3, 417–424.

Searle, J. R. (1992). *The Rediscovery of the Mind*, Cambridge, MIT.

Severson, R. L., and Carlson, S. M. (2010). Behaving as or behaving as if? Children's conceptions of personified robots and the emergence of a new ontological category. *Neural Networks*, 23, 1099–1103

Suppe, F. (1989) *The Semantic Conception of Theories and Scientific Realism* University of Illinois

Tapus, A., Mataric, M., Scassellatti, B. (2007). The Grand Challenges in Socially Assistive Robotics. *IEEE Robotics and Automation Magazine*, Institute of Electrical and Electronics Engineers.

Turkle, S. (2010). In good company?, in Wilks Y. (Ed) *Close Engagements with Artificial Companions*, Benjamins, 3–10

Webb, B. (2001). Can robots make good models of biological behavior? *Behavioral and Brain Sciences*, 24, 1033-1050.

Wolpert, L. (1992) *The Unnatural Nature of Science*, Harvard University Press.

Zawieska K and Duffy BR (2014) The self in the machine. Pomiary, Automatyka, Robotyka 18(2), 78-82.