The robots pictured in this gallery, all of them artificial agents developed by external social robotics, are ordered from beginning to end along the Mori curve. The robots that most closely resemble humans are located at the end of the gallery, and those that resemble them least at the beginning. This distribution conceals what, strictly speaking, should be regarded as an ambiguity inherent in the concept of resemblance. The criterion employed here is visual resemblance. According to another criterion, affective expression, the ranking is different. Keepon's abilities, for example, are far superior to those of Saya. Even if visually Keepon resembles us much less than Saya does, it is much more similar to us emotionally. This ambiguity makes it possible to detect the various ways in which roboticists exploit people's instinctive anthropomorphism to draw the user into an affective and empathetic relationship that allows the robot to effectively perform its role of social or therapeutic mediation. The multiplicity of resemblances also lends support to the new view of anthropomorphism as an essential element of human cognitive abilities.

Internal Robotics, or The Private Dimension of Emotion and Artificial Empathy

Research on the internal robotics of emotion is concerned with the role of affect in organizing behavior. In fields such as cognitive robotics, epigenetic robotics, and developmental robotics, the objective is twofold. First, researchers attempt to design and physically implement artificial models of natural affective processes to explore experimentally how emotions contribute to cognition and purposeful activity. The motivation of basic research, in other words, is to arrive at a better understanding of emotion in human and animal behavior. Second, researchers try to use this knowledge to build robots having greater autonomy and an improved capacity for adaptation. Implementation of systems for regulating emotions in robotic agents is meant to equip them with mechanisms that allow them to unilaterally establish an order of priority among various behavioral options. These robots will be capable of selecting appropriate learning and adaptation strategies in response to the needs of the moment, which in turn will increase their capacity for autonomous action with the environment and with other agents, robotic and human.

The growing attention to the internal aspect of emotion is part of a paradigm shift presently taking place in cognitive science, away from the old computational model toward the conception of embodied mind we discussed earlier. This new paradigm, which puts the body back at the center of the scientific study of cognitive processes, has revived interest in phenomena that were neglected by classical cognitive science because they were considered to be merely corporeal, devoid of cognitive value.[35] In robotics, the importance now attached to emotion in cognitive processes—a consequence of the increasingly accepted view that the two are in fact indissociable[36]—has led to the development of radical versions of the embodied mind

thesis and, in particular, to an attempt to go beyond the doctrine of sensorimotor functionalism, which seats cognitive architecture in the perception-action cycle. One of the most promising new directions for research, known as organismic embodiment,[37] seeks to recreate within artificial systems the complex interrelations between cognitive processes and affective regulation that are characteristic of natural cognition.

From the methodological point of view, the development of "embodied cognitive science"[38] has led to a gradual abandonment of the sort of representational modeling of cognitive phenomena typically done in classical artificial intelligence,[39] where they are treated as essentially or exclusively computational, in favor of so-called constructive or synthetic modeling.[40] The new approach incorporates hypotheses about cognition in robotic systems and tests them experimentally by analyzing the behavior that these systems express in appropriate environments.[41] Here one of the operative assumptions is that cognition is expressed directly in behavior, which is no longer conceived simply as an indirect manifestation of a system's cognitive capacities.

This constructive approach has a long genealogy in the sciences of self-organization.[42] In its current version, it constitutes a form of explanation that, contrary to the method long dominant in modern science, seeks not to reveal the underlying simplicity of complex phenomena by reducing them to the simple mechanisms that generate them, but to recreate complex phenomena in all their complexity by reproducing the dynamic interrelation of organizational levels: the system, its elements, the environment—and all of their interactions, not least the ones involving the observer.[43]

A paradigmatic example of this approach, applied to the modeling of emotions, may be found in an essay published some thirty years ago by Valentino Braitenberg.[44] The artificial agents imagined by

Braitenberg—vehicles that move around in space, avoiding obstacles, moving away from one another or away from light sources, or moving closer to them, and so on—have a simple architecture that contains only sensorimotor correlations. Yet when these agents interact with the relatively rudimentary environment that their perceptual abilities allow them to recognize, they display behaviors that an observer would spontaneously describe in terms of emotions: they are afraid, they have desires, they are aggressive. This modeling has the advantage of bringing out in a very clear fashion the interactive character of the synthetic approach and the indispensable role of the observer. It has the defect, however, of limiting itself to an abstract sensorimotor functionalism and of applying the synthetic method to no useful purpose.[45] For all its virtues, Braitenberg's thought experiment remains just that, a thought experiment: the cognitive architecture he describes has never been realized, and it lacks any structure or dynamic capable of producing emotions in the way they are produced in natural systems. Affective phenomena in Braitenberg's modeling are open to the same objection as the external robotics of emotion, namely, that they exist solely in the mind of the observer. The synthetic method, as it is used here, is incapable of testing hypotheses about how emotions are produced or of exploring experimentally the role they play in organizing behavior.

Studies on the genesis and functioning of affective processes now being carried out by researchers in cognitive robotics, epigenetic robotics, and developmental robotics under the organismic embodiment paradigm use the synthetic approach in a more productive way. Unlike the classic representational approach of cognitive-affective robotics, which consists simply in using "boxes, standing in for ad hoc mechanisms, that *label* states as 'emotions,' 'feelings,' etc.,"[46] organismic embodiment seeks to use "mechanisms that are argued to be constitutive of representative and / or emotional phenomena,

[an] approach [that] offers greater scope for emergence and flexibility in robot behavioral performance."[47] These mechanisms are implemented in a real robotic architecture, rather than evoked in a purely conceptual fashion in the description of processes supposed to underlie affective behaviors. This makes it possible, in principle, for the observer to wholly recast the dynamics of emotional phenomena within an experimental framework of interactions between a robotic agent and its environment.

It must be emphasized that this research is only in its early stages, and that the promise of the constructive approach remains for the moment unfulfilled. In a striking passage of his recent programmatic manifesto, Domenico Parisi describes the challenge facing researchers today:

> The brain of our robots is extremely simple and it should be made progressively more complex so that its structure and its functioning more closely match what we know about motivations and emotions in the real brain. But the brain is not enough. Motivations and emotions are not *in* the brain. They are the result of the interactions *between* the brain and the rest of the body. The emotional neurons of our robots should influence and be influenced by specific organs and systems inside their body—the equivalent of the heart, gut, lungs, and the endocrine and immune systems. But this is a task for the future. . . . [I]f robots must reproduce the behaviour of animals and human beings and, more specifically, their motivations and emotions, what is needed is also an *internal* robotics.[48]

Parisi's call to arms has not been ignored. Already work is under way aimed at building abstract models of robotic agents.[49] In the meantime, efforts to construct and implement robotic platforms are continuing,[50] and new directions for research are beginning to be explored as well.[51] Unlike attempts in the external robotics of emotion to awaken and exploit our instinctive anthropomorphism, these

studies are devoted to creating emotions and empathy in robots. The artificial agents that internal robotics hopes to fabricate will be genuinely empathic, and their emotions real, because authentic. As the cognitive neuroscientist Ralph Adolphs puts it:

> [R]obots could certainly interact socially with humans within a restricted domain (they already do), but . . . correctly attribut[ing] emotions and feelings to them would require that robots are situated in the world and constituted internally in respects that are relevantly similar to humans. In particular, if robotics is to be a science that can actually tell us something new about what emotions are, we need to engineer an internal processing architecture that goes beyond merely fooling humans into judging that the robot has emotions.[52]

The internal robotics of emotion aims therefore at one day producing artificial agents that may be considered emotionally and socially intelligent. The hypothesis guiding this research is that these agents will display an intelligence similar to ours to the extent that their affective and emotional capabilities rest on an internal architecture constructed on the basis of deep models of human social skills (see Table 1).[53]

In seeking to create robots with true, sincere, authentic emotions, the internal approach counts on being able to deflect the charge brought against the artificial agents of external robotics—that their emotions are false, because feigned—by equipping its agents with internal mechanisms that are supposed to play the same role as emotions in regulating human behavior. The principal limitation of this approach is not so much that it reduces emotion to a particular functional role as that, once again, it confines affect to a person's relationship to himself.

TABLE 1

The Internal Robotics of Emotion

Approach	Description
Neuronal Network Model[a]	Based on the dopamine system of the mammalian brain
	Implemented on a simple robotic platform, MONAD
Cognitive-Affective Architecture[b]	Based on three different levels of internal homeostatic regulation and a form of behavioral organization that integrates neural and corporeal activity and sensorimotor activity
	Has different levels of organization that are directly linked to behavioral tasks and to the robot's autonomy
Developmental Affective Robotics[c]	Recursive approach at several levels that takes into account and interconnects the various phases of human physiological and psychological development
	Rests on the principles of cognitive developmental robotics; centered on a knowledge of self and others
	Asserts a parallelism between empathetic development and the development of self / other cognition

a. See W. H. Alexander and O. Sporns, "An embodied model of learning, plasticity, and reward," *Adaptive Behavior* 10, nos. 3–4 (2002): 143–159.

b. See Anthony F. Morse, Robert Lowe, and Tom Ziemke, "Towards an Enactive Cognitive Architecture," in *Proceedings of the International Conference on Cognitive Systems [CogSys 2008]*, Karlsruhe, Germany, April 2008; www.ziemke.org/morse-lowe-ziemke-cogsys2008/.

c. See M. Asada, "Towards artificial empathy: How can artificial empathy follow the developmental pathway of natural empathy?" *International Journal of Social Robotics* 7, no. 1 (2015): 19–33.

Constructing an Affective Loop

The fundamental question for the future of research in the robotics of emotion has to do with the relation between its internal and external aspects. Recent developments suggest that the customary distinction between the two is in fact no longer a settled matter and that social robotics, in upholding the distinction while at the same time disregarding it, harbors a deep ambivalence in this regard.

It upholds the distinction to the extent that the external and internal aspects tend to be treated separately and from different theoretical perspectives. Specialists characterize this divide by means of a series of oppositions, between the social and the individual, the interindividual and the intraindividual, and, above all, between false (or feigned) emotions and true (or authentic) emotions. And yet no justification is given in the current robotics literature for the distinction itself, nor for the series of oppositions that is supposed to correspond to it. Why should the internal aspects of emotion be true and the external aspects false? The question seems never to have been posed. The distinction and corresponding oppositions are simply assumed to go without saying. But in fact they all derive from a commonsensical conception of emotions, central in philosophy at least since the nineteenth century[54] and subsequently taken over not only by classical cognitive science, but also by moderate variants of the embodied mind approach, dominant today.

According to this conception, emotions are internal, essentially private events that take place in an intraindividual space. They are directly accessible to the subject of affective experience and subsequently may or may not be, depending on the case, publicly expressed. When they are, they become accessible to others, but indirectly through the fact of their expression, which remains contingent. Intersubjective knowledge of emotions, just like knowledge of others' minds since Descartes, is therefore never direct. It is the result of

rational analysis of the behavior of others, and it rests on an analogy with—which is to say that it is worked out on the basis of—what the subject himself feels and his own reactions in similar circumstances.

It is this instinctive view that leads roboticists to divide their studies into two parts, one bearing on the internal aspects, the other on the external aspects of emotion and empathy—a division that is supposed to correspond to the boundary separating intraindividual space from extraindividual space. In practice, however, it does not really guide research in either one of these two approaches. Each one denies in its own way the distinction that forms the basis for the difference between them; both agree on one crucial point, namely, that affective phenomena occur in a space that encompasses both the intraindividual *and* the extraindividual.

Internal robotics, which claims to be devoted exclusively to modeling the internal machinery of emotion, proceeds, as we have seen, by means of a synthetic or constructive method. Now, this method is inevitably opposed, at least up to a certain point, to the classical thesis that emotions are private events that are generated in intraindividual space and contingently expressed in extraindividual space. The synthetic method in social robotics seeks an explanation of the phenomena it studies in the joint relationship between an agent and the environment in which the agent operates. The relevant unit of analysis in current synthetic modeling is therefore not the isolated individual, but the system constituted by the agent and its milieu—a unit that is inherently relational and, within the framework of social robotics, indisputably characterized as *inter*individual.

The external robotics of emotion likewise, and more obviously, adopts a relational approach to affective phenomena. It, too, violates the distinction between internal and external, only now from the other direction. Even when they are explicitly inspired by the classical view of emotion, studies in external robotics assign the expression of emotion a role that is not limited merely to communicating

predefined feelings to a human partner. The affective expression of the artificial agent is conceived and modeled as playing an active part in the genesis of human emotions; in other words, even though the robot's affective expression is not associated with any inner experience or internal regulatory dynamic, just the same it is explicitly intended to produce an emotional response in its human interlocutors. Expression is therefore central to the processes that generate emotions, that is, the affective reactions of human interlocutors, which nonetheless continue to be thought of as inner experiences. Plainly, then, the robot's emotive expression transgresses the strict separation of internal from external, of the individual from the social.

It is plain, too, that the affective reaction of human partners does not rely on any understanding of what the robot feels, whether "by analogy" or in any other way. We know very well that robots feel nothing, and while the physical appearance of Geminoids or of Saya may fool us occasionally, NAO and Keepon are manifestly robots—and this does not prevent them from stirring our own emotions. Even if our emotions are generally not considered to be authentic in this case, because they have been produced by a robot's feigned emotions, which deceive us, the truth of the matter is that this value judgment tries to discredit what it cannot help but confirm, namely, that our emotional response is in fact quite real.

Indeed, in the case of the artificial empathy contrived by external robotics, we are dealing with *pure affective expression.* There is no interiority here, no corresponding inner states. Instead, what we have is a dynamic that does not take place inside, in intraindividual space; it occurs wholly within the space of a *robot-human relational unit.* This interactionist perspective permits us to recognize external robotics as a synthetic approach to modeling emotions and empathy, a minimalist application of the synthetic method that in certain respects resembles the one imagined by Braitenberg's thought experiment.[55] Both rely on a rudimentary cognitive architecture that, in interaction

with the environment, allows emotions to be perceived by an observer. By comparison with Braitenberg's synthetic psychology, however, external robotics exhibits two fundamental and closely related characteristics. First, it is more than a mere thought experiment. Second, it creates a real affective and empathetic response in a human being, the robot's interlocutor, who now can no longer be regarded simply as an external observer of an artificial system. The whole point of external robotics consists in just this, that when the interlocutor observes or perceives affective expression in the robot, he immediately enters into the affective process and actively takes part in sustaining the dynamic that drives it.

The boundary between inward and outward aspects of emotions and empathy, which is supposed to coincide with the limits of *intra*individual space, is continually overstepped as much by internal as by external robotics. They are both agreed, in spite of their divergent orientations, in situating affective processes not in the individual agent, but in the relationship between the agent and his environment, in the *inter*individual unit formed by agent and environment.

It hardly comes as a surprise, then, that an interactionist perspective concerned with reconciling the internal and external approaches should begin to emerge in social robotics today. In seeking to create an "affective loop,"[56] it operates on the assumption that robots must be capable of including the people who deal with them in a dynamic that encompasses the poles of the relationship (robot and human interlocutor), affective expression, and the responses that such expression evokes. The process is interactive, with interaction taking the form of a loop. The essential thing is that the system's operation "affects users, making them respond and, step by step, feel more and more involved with the system."[57] The way to ensure that this will happen is to improve the robot's *social presence,* which in turn means that "social robots must not only convey believable affective

expressions but [must] also do so in an intelligent and personalized manner."[58]

Situated at the intersection of internal and external robotics, this interactive approach looks to combine the two dimensions of emotion, both of which are recognized to be necessary to generate a functional affective loop. This requires that the robots be equipped with an internal affective architecture and an external capacity for emotional expression. The degree of complexity of the different internal architectures that may be implemented depends on the way in which they are modeled on natural processes responsible for producing emotions and also on how great a resemblance between natural and artificial processes is thought to be desirable. As in the case of natural systems, expressive abilities are associated with the internal architecture in a variety of ways. A few preliminary examples of robots produced by the interactionist approach are shown in Gallery 2.

First, there are robots like BARTHOC Jr. (a smaller twin of the Bielefeld Anthropomorphic RoboT for Human-Oriented Communication) that, according to Breazeal's classification, are situated along the boundary between "social interface" robots and "socially receptive" robots.[59] Even if they are socially passive, they are nevertheless capable of benefiting from human contact. An interactive human-robot interface is implemented through the imitation of the human partner's observed emotions, which gives rise in the robot to a "simple" model of human social competence. Next, there are robots such as Kismet, which Breazeal calls "sociable robots."[60] These artificial agents, no matter that they are still far from having the desired form of intelligence, are nonetheless capable of including their interlocutors in a minimal affective loop. This permits the robot to attain what Breazeal calls internal subjective objectives. In other words, the minimal affective loop that is established between artificial and human agents allows the robot to acquire affective states that are appropriate to the

current stage of their evolving relationship and that are determined on the basis of an internal model of social knowledge and emotional intelligence.

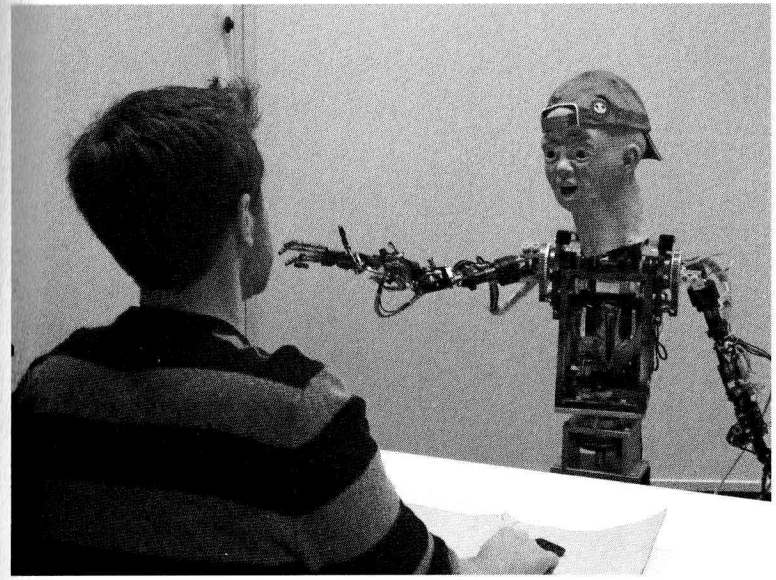
The aim is to give robots the means to construct an increasingly stable and efficient affective loop. One way of doing this, currently being studied, is to equip an artificial agent with its own biography, supplemented by a "self-narrating" ability that permits it to provide interlocutors with a "personal" interpretation of its life. The robot therefore has a first-person memory that contains a record of individual experience on which it may draw in communicating with human partners. In this fashion, it is hoped, it will be able to acquire a more consistent and more convincing social presence, similar to the presence of *someone,* to the presence of a person.

The interactionist approach is therefore intended to reduce and eventually eliminate the contrast, and even the distinction, between true human emotions and false robotic emotions. Even if the internal models that produce affective behavior have nothing like the "thickness" of the processes modeled by human or animal physiology, it would no longer be possible to consider robotic emotions as merely feigned or false, for these emotions represent and reveal a genuine inner event. They would no longer be simply a means of fooling us, of making us think that the robot has an emotion it does not really have. Over time, then, it is hoped that the affective expressions of artificial agents, because they are now correlated with an internal dynamics, as in the case of human emotions, will become gradually truer, more authentic.

Note, however, that the attempt to give robots genuine emotions has the unexpected effect of reconciling the interactive approach with the classical conception of the emotions. Interaction, however real it may be, remains external to emotion, which in this conception of an affective loop effectively constitutes an inner state. It is the "inner state" of a robot that tells us the truth about its emotions.

GALLERY 2

The Interactionist Approach



BARTHOC Jr.

Appearance: Partially human-like

Emotional / Empathetic Interaction: The robot recognizes certain emotions (joy, fear, neutral affective state) of its users through the analysis of spoken language. It expresses ("mirrors") these emotions by its facial expressions.

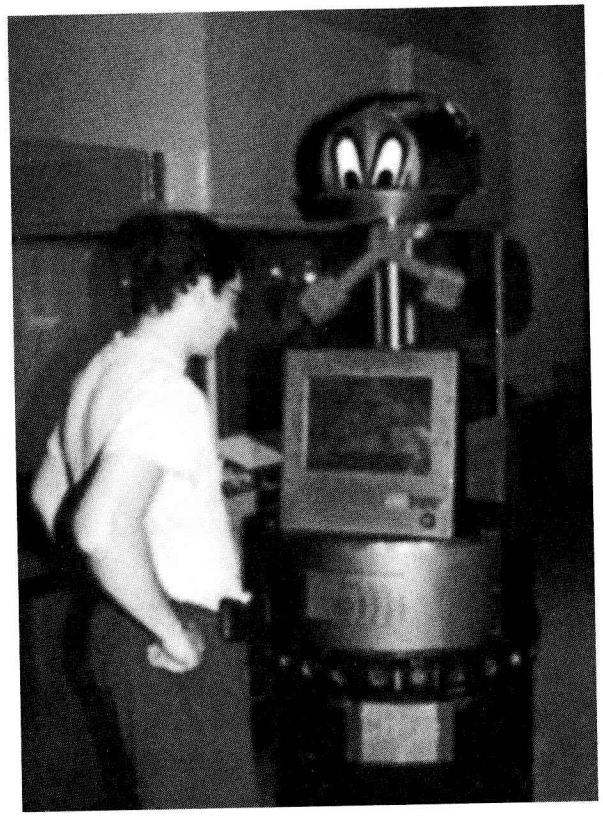
Principal Use: Exploring human-robot interaction

Maggie

Appearance: Cartoon-like

Emotional / Empathetic Interaction: The robot is equipped with an emotional control system whose objective is to maintain the interlocutor's well-being. When it perceives a change in an indicator, from greater well-being up to a certain affective threshold (joy, rage, fear, or sadness), it becomes active; the robot's decision-making system determines the robot's action on the basis of (1) emotional motivations, or drives, (2) self-learning, and (3) its current state. The robot utilizes facial expressions as well as movements of the body, arms, and eyelids.

Principal Use: Exploring human-robot interaction; helping others (as a nurse's aide, for example); entertainment

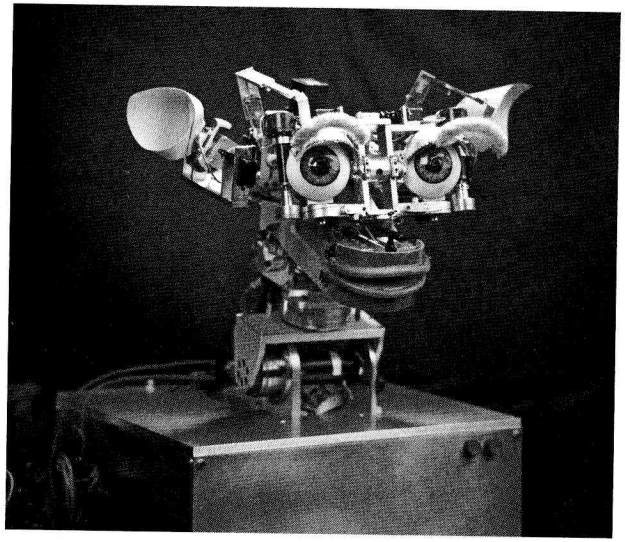


Sage

Appearance: Cartoon-like

Emotional / Empathetic Interaction: The robot's cognitive architecture is capable of giving it an affective "personality" through changes in its moods: happy, occupied, tired, lonely, frustrated, confused. Certain predefined events gradually lead to these mood changes. The robot's feedback to its interlocutors consists in expressing its moods through speech, the tone and register of voice, and the volume and speed of speech as well as what is said.

Principal Use: Instruction (as a museum guide, for example)



Kismet

Appearance: Cartoon-like

Emotional / Empathetic Interaction: The robot's actions are commanded by an affect-recognition system, an emotion system, a motivation system, and an expression system. The emotion system is inspired by an ethological model for perception, motivation, and behavior; emotions are triggered by events having significant consequences for the robot's "well-being." When an emotion is activated, it leads the robot to enter into contact with something that promotes its well-being and to avoid anything that is contrary to it. The affect-recognition system influences the robot's facial expression; on the basis of these expressions, the human partner can interpret the robot's affective state and modify his behavior and interaction accordingly. The robot's face can express anger, fatigue, fear, disgust, excitement, happiness, interest, sadness, and surprise.

Principal Use: Exploring human-robot interaction; therapeutic mediator for autistic children

To be sure, we are capable of constructing such states, but we will never have direct experience of them, neither more nor less than we will have direct experience of the emotions of another person. In each case, we are supposed to be satisfied with knowing that they are there—that they exist.

The truth of an emotion remains foreign to affective interaction. Evidently, the classical paradigm has not quite yet been abandoned after all.