the subject from sharing his point of view, but that subjective error can only be conceived or imagined in relation to another agent having a different point of view. This agent alone makes possible the hyperbolic doubt on which the certainty of one's own existence depends, the assumed epistemic privilege of first-person knowledge of the mind. Hyperbolic doubt requires the existence of another cognitive agent who can take me as an object of knowledge and declare that I am mistaken.

Descartes's hyperbolic doubt and the metaphysical discovery of the mind to which it gives rise require the presence of someone who is able to judge that the subject, Descartes, is mistaken. What permits, though evidently it does not justify, the claim that knowledge of oneself and of one's own mind enjoys epistemic priority over one's knowledge of the world, particularly knowledge of others, is the active presence of a second epistemic agent who exercises his cognitive abilities in the same world in which Descartes exercises his own. For the argument to be successful, for Descartes to be convinced of the certainty of his own existence, Descartes himself must be taken as an object, must be operated on, so to speak, by another epistemic agent who will have "devoted all his energies" to deceiving him. The epistemic priority of first-person knowledge is an illusion, for both the discovery of the mind and the impression that it occupies the center of the cognitive process are made possible by, and flow from, the presence and the action of another epistemic agent who acts upon the subject.

The environment from which the characteristics of the human mind emerge—characteristics that, since Descartes, have been supposed to define its superiority by comparison with other cognitive systems—is a social environment. The cardinal virtue of the Cartesian metaphysical fable is that it implies just this, that *the subjective structure of first-person experience of the mind does not reflect the structure of the cognitive process from which it emerges.* To be sure, Descartes himself

drew an altogether different conclusion. Ever since the publication of the *Meditations,* the epistemic priority of knowledge of oneself and the mind has generally been accepted as something obvious. A close reading of what Descartes actually says makes it clear, however, that the discovery of the mind is the result of being fooled—of being fooled, or otherwise led into error, by another epistemic agent. The social environment inhabited by the Cartesian subject is therefore not emotionally neutral. It is shot through with affect, riddled with the anxiety of a subject who imagines not only that he has been led into error by another, but that the error is one that he is not even capable of conceiving.

Methodological solipsism and the priority of self-knowledge over knowledge of others therefore have no sound philosophical basis. There is no reason to take at face value the spontaneous and commonsense answer to the question "Where is the mind?"—neither in its usual version, nor its metaphysical version, nor in the vague and underdetermined version proposed by the extended mind thesis. The mind is neither in the brain, nor in the head, nor outside the agent (much less in Otto's address book), *but in the relations that obtain between epistemic agents.* This is why the question "Where is the mind?"—just like the question "How far does the mind extend?"—scarcely makes any sense at all.

Emotive and Empathic Robots

The branch of robotics that explores the social environment from which mind emerges, a domain largely neglected by philosophy of mind and cognitive science, assigns a central place to the study of affect. Compared with other disciplines, the mixed empirical / theoretical style of research practiced by social robotics gives it a distinctive flavor. On the one hand, it seeks to devise solutions to engineering problems, so that robots will have this or that functional property or

skill; on the other hand, it is guided by a generally shared understanding of sociability and the emotions. There is a tension, however, even a contradiction, between these two aspects. Considering emotions from the theoretical point of view, social robotics adopts the methodological solipsism dominant in philosophy of mind, cognitive science, and psychology, whereas in practice, which is to say for purposes of research and technological applications, affect is seen to be a social phenomenon.

Social robotics is an extremely dynamic field of interdisciplinary inquiry situated at the intersection of a number of different research domains, among them human-robot interaction,[10] affective robotics,[11] cognitive robotics,[12] developmental robotics,[13] epigenetic robotics,[14] assistive robotics,[15] and rehabilitation robotics.[16] They encompass a variety of theoretical frameworks, objectives, and modes of inquiry that are constantly changing, not only in the case of a particular research project, but even within a particular team of researchers. All of these perspectives nonetheless converge and contribute to social robotics through their shared interest in a single fundamental question: how can artificial agents of a new type, social robots, be introduced into our network of social interactions? These agents are not designed to function simply as robotic servants in public settings associated with information, education, health care, therapeutic mediation, entertainment, and so on. The term "social robots" refers more precisely to artificial agents that are able to work in these fields by virtue of their ability to *socially* interact with human beings.[17]

The defining characteristic of social robots, as we saw earlier, is the ability to be perceived by those who interact with them as being *present* in the way a person is. To exhibit social presence, in other words, a robot must give its human partner the "feeling of being in the company of *someone,*"[18] of being confronted with an "other." This is not a mere act of projection on the part of the human partner, but a matter of actually causing him to feel something that is a crucial element of face-to-face encounters, the basic relationship from which, in the last analysis, all other social relationships are derived.

Social presence involves something more, and something other, than the ability to decode verbal messages and to respond to them appropriately, or to identify other agents and recognize them in different situations. These high-level cognitive abilities present difficulties in implementation that classical computational artificial intelligence has been working to resolve since its inception. Social robotics studies other kinds of interaction as well, foremost among them the communication that is created by means of what is popularly called body language—posture, gesture, gaze, physical contact. This form of communication also includes affective reactions,[19] in which felt presence is left over as a sort of precipitate. The hope of present-day researchers in social robotics is that by taking into account these various physical and relational elements, which go beyond and at the same time modify the purely computational aspect of the relationship with human partners, it will be possible to find simpler and more satisfactory solutions to problems that have so far proved to be, if not quite intractable, nonetheless very difficult.

The importance of the affective dimension of human-robot relations, especially in helping promote the "social acceptance" of robots,[20] has focused attention on the challenge of building robots that can recognize and correctly interpret the emotional manifestations of their human partners and respond to them appropriately. Indeed, the ability to sustain empathetic relations with human interlocutors is now taken to be the fundamental criterion of successful behavior by robots in social contexts.

The construction of "affective," "emotive," or "empathic" robots, as they are variously called, is currently at the forefront of research in social robotics.[21] Our view is that this is the most promising work now being done, not only in social robotics proper, but also in cognitive science as a whole. The robotic implementation of emotions and

empathy has implications that go beyond creating artificial agents endowed with social presence. Apparently straightforward attempts to give emotions and empathy a positive role in robot interactions with human beings inevitably involve a variety of problems at the boundary between cognitive science and philosophy of mind that can then be investigated experimentally. These developments converge with the argument we have developed so far to the extent that social robotics calls into question the fundamental limitation imposed upon cognitive agents by classical cognitive science and philosophy of knowledge: the prison of the solitary mind in which they have been locked away.

In particular, research in social robotics challenges the customary use of singular possessive pronouns in expressions such as "*my* mind" or "*your* mind"[22] and urges us to abandon the assumption, inherited from traditional philosophy, that the mind is essentially something internal, individual, and private, of which the agent is, as it were, the owner. This assumption is common to all the main schools of thought in cognitive science, including, as we have seen, the emerging conception of the mind as a spatial entity, situated in the brain, that may sometimes extend beyond the confines of skull and skin. This view, which purports to be revolutionary for contradicting (or so it imagines) the internalist thesis formulated originally by Descartes, in fact does nothing more than cloak the underlying Cartesian dualism of mind and matter in a fashionable physicalist monism.

The development of a robotics of emotion suggests, to the contrary, that mind is neither immaterial nor something extended in space, but a network—better still, to recall the definition made famous by Gregory Bateson, an "ecology."[23] We have already seen in Chapter 2, and will presently see in greater detail, that mind can be conceived as the coupling or interconnection of different cognitive systems among themselves and with their environment, as well as of

the multiple modes of dynamic coordination that link them together on different levels.

Yet the influence of social robotics in both illustrating and helping bring about a paradigm shift in how we think about the mind is not limited to academic debates in cognitive science and philosophy. The way in which emotions and empathy are treated by social robotics has a direct bearing as well on our understanding of the very considerable impact it is likely to have on daily life in our social ecologies. Considering the explosive growth this field of research is expected to undergo in the years ahead, it is scarcely possible to overstate the importance of taking a close look now at what the future holds in store for us.

A Vanishing Divide

Dualism dies hard, and even in social robotics traces of it can still be found. Here two major approaches have long been dominant, one concentrating on the "external" aspects of emotion, the other on its "internal" aspects. These two approaches are associated respectively with the social and individual dimensions of emotion.[24] Research has consequently followed separate paths, one for each domain, even though everyone recognizes that these dimensions are closely related and, what is more, that they have coevolved.[25] But while the distinction between the two approaches is steadfastly upheld within the research community, in practice the difference between them has proved to be unstable and uncertain, and as we will see, many recent developments that are generally supposed to provide support for it have had the effect instead of undermining it further. This state of affairs has given rise to attempts to connect and integrate the internal and external aspects of emotion, as well as the individual and social dimensions to which they correspond. In Chapter 4, we evaluate the

success of these efforts. It will be necessary, we believe, to try to go still farther and to reject altogether the dualism underlying this very distinction. Rather than try to join together these two dimensions as separate facets of emotion, we need to conceive of them as integral moments of a continuous dynamic.

The distinction between the two basic orientations of research in social robotics reflects two paramount preoccupations. One involves affective expression by robotic agents, regarded as an external phenomenon; the other is concerned with the production and regulation of emotions in robotic agents, regarded as an internal phenomenon. From the technical point of view, the two orientations assume different forms. Research on the external (or social) aspect of emotions seeks to devise robots that exploit our propensity to anthropomorphism in order to provoke emotive and empathetic reactions in users, whereas research on the internal (or individual) aspects aims at constructing robotic agents whose behavior is influenced by a form of affective regulation inspired by the natural regulation of emotions in human beings and animals.

Accordingly, the distinction between external and internal approaches is associated with two quite different views of robotic emotions. In the first case, the emotions displayed by robots, to the extent that they are *merely* expressed, are thought to be feigned and "false." Exclusive attention to affective expression is bound to produce machines that are limited to simulating emotion—robots that may sometimes be capable of reacting to the feelings of human partners in a suitable and productive fashion but that themselves do not actually feel anything. Such artificial agents are therefore open to the objection that they fool us; that they function by means of deceit, for they do no more than take advantage of the credulity of their human partners, who, moreover, are apt to be vulnerable, as is the case with children with special needs and the elderly.

In the second case, by contrast, the point of equipping robots with a biologically inspired capacity for affective regulation is to create artificial agents having "real" robotic emotions, whose functional role is similar to that of human or animal emotions. The idea, in other words, is to build robots that do not deceive, robots that actually have emotions—"true" emotions, whose expression would be genuine. We are still far from being able to do this, of course. But the basic assumption guiding current research is that artificially reproducing the functional role of emotions is necessary, and perhaps also sufficient, if robots are to be endowed with "real" emotions and that this constitutes an indispensable condition of their entering into authentic affective relationships with human beings.

Beneath these two approaches, and the distinction between the two complementary aspects of emotion that they enforce, we find the same conceptual structure as before: a private mind (in the event, private emotions) shut up inside an agent that can only be known by another mind indirectly, on the basis of the agent's outward behavior. The trick of external robotics—or the lie, as some may think of it—therefore consists in its mimicking these external manifestations without there being anything inside the agent, any inner feeling underlying them. Internal robotics proposes instead to provide the artificial agent with a form of interiority that serves to guarantee the authenticity of the emotions it displays. There is no escape from the dualist schema in either case, however, even if the "interiority" that is in question here, whether present or absent, could not possibly be more materialist. The moral issue connected with the distinction between these two approaches in social robotics, which is inseparable from the question of what constitutes a true and undeceitful affective relationship, is unavoidably affected by the same dualism.

But the habit of associating the external approach with false emotions and the internal approach with true emotions turns out, on

closer examination, to be without foundation. Current developments in social robotics are steadily eroding the demarcation between internal and external aspects of emotion, undermining the idea that real, unfeigned emotions—genuine emotions—necessarily arise from a strictly internal process that constitutes the guarantee of authentic affective expression. This challenge to the traditional external / internal division will lead us to reformulate the moral question and approach it from a fresh perspective.

External Robotics, or The Social Dimension of Emotion and Artificial Empathy

Research today on the external or social aspects of emotion focuses on the expressivity that movement, gesture, posture, and proxemic aspects, as well as body type and face shape, give to robots. The purpose of studying the influence of the static or dynamic appearance of robots on interactions with human beings is to equip artificial agents designed to perform a variety of services with forms of expressivity that, by giving emotional color to their actions, will facilitate their acceptance and increase their effectiveness. At the heart of this approach is the idea that empathy plays an essential role in establishing convincing, positive, and lasting affective relationships between robots and humans.

The attempt to build empathic robots is a highly interdisciplinary enterprise, which involves a two-way transmission of knowledge between very different domains, from the performing arts to cognitive psychology and the natural sciences. Basic research is concerned chiefly with the production and recognition of affective expressions, paying particular attention to the factors that favor anthropomorphism, which is to say the spontaneous tendency to attribute beliefs, intentions, desires, emotions—in short, mental states—to animals and to a wide range of artifacts, from stuffed animals to androids. The scientific

conception of anthropomorphism has been profoundly transformed and revitalized by recent work in cognitive science and emerging fields of research that come directly under the head of social robotics or that are closely associated with it, such as human-robot interaction and human-computer interaction, with the result that new ways of developing an external robotics of emotion can now be explored.

Psychologists have traditionally conceived of anthropomorphism as the result of confusing the physical and the mental. According to Jean Piaget, this confusion is characteristic of the "egocentric" and "animistic" thinking that children commonly exhibit until the end of the seventh year.[26] More recent studies consider, to the contrary, that anthropomorphism constitutes a fundamental dimension of the human mind. In this view, anthropomorphism is not limited to a particular phase of development; it is *independent of the beliefs of agents as to whether the objects with which they interact have mental properties.* The propensity to attribute mental states to inanimate objects is nonetheless *strongly influenced by the nature and context of the interaction.*[27]

The central hypothesis of this new conception is that our actions take place and our thinking evolves principally by means of dialogue, which creates a context in which we spontaneously treat animals and artifacts as interlocutors. A dialogue context is defined as any communication situation in which turn taking is likely to take place. Such situations can be created in many different ways, through imitation, verbal expression, nonverbal vocal expression (cries, grunts, groans, and so forth), and gestural expression. Currently, there is a consensus among researchers, robustly supported by experimental results, that anthropomorphism derives from the operation of fundamental cognitive structures, which is related to our tendency to think teleologically and to interact through dialogue. The activation of these structures would explain why anthropomorphic projections occur in relationships where the content of the interaction is very poor, and with the full awareness that the interlocutor, typically

an animal or an artifact, does not in fact possess the properties that we attribute to it during the interaction. When we plead with a computer not to break down just now ("This *really* isn't a good time, you know!"), we do not suppose that it can hear us or understand us, nor do we imagine that our entreaties will have any effect whatever.

This way of looking at anthropomorphism suggests that feeling comfortable in the presence of an artifact and developing a shared sense of empathy depends on certain minimal conditions that allow it to be treated as an interlocutor in the first place. The external approach seeks to satisfy these conditions through forms of embodiment and autonomous movement that give robots dialogical skills ranging from simulation—gestural and facial reactions suggesting an interactive presence—to actual conversational abilities, leading to the production of suitable verbal responses. The fact that these capabilities can be realized in several different ways makes robots valuable instruments for research not only on anthropomorphism itself, but also on the behavior of artificial agents in various social contexts, whether they are employed as therapeutic aides, teaching assistants, or receptionists, or in any other capacity where their talent for arousing affective and empathetic reactions may prove to be useful.

Throughout this broad field of investigation, in which pure research is conducted alongside the manufacture of technological devices, one of the important sources of inspiration is Masahiro Mori's "uncanny valley" conjecture, which we considered in Chapter 1.[28] To create an impression of familiarity, a robot must resemble human beings in a number of crucial ways, but it must not be *too much like us.* Mori postulated, without really knowing why it should be, that too great a resemblance will give way to a sense of unease, discomfort, anxiety, and, in the extreme case, revulsion. Several recent studies in the field of human-robot interaction have shown that resemblance is not, in fact, decisive in determining how comfortable we feel in the

company of robots.[29] In agreement with the revised interpretation of anthropomorphism, these studies indicate that the attribution of emotional and empathic properties to robotic agents depends mainly on the specific characteristics of a given interaction. They agree, too, with the explanation for the sudden collapse of familiarity represented by descent into the uncanny valley that we advanced earlier—an explanation that at once deconstructs Mori's conjecture and accounts for the phenomenon he sought to describe.[30]

All the robots in Gallery 1 fall within the first two categories in the classification of social robots drawn up by Cynthia Breazeal.[31] To begin with, there is a class of *socially evocative robots* that "rely on the human tendency to anthropomorphize and capitalize on feelings evoked when humans nurture, care [for], or are involved [with] their 'creation.'"[32] Next, there is a class of *socially communicative robots.* These offer a natural communication interface to the extent that their capacity for social communication, which rests on a "shallow" model of human social cognition, is nonetheless sufficiently similar to what we are capable of. Breazeal believes that the external approach in social robotics cannot hope to go any further than this and so will be unable to produce artificial agents corresponding to the higher levels of her classification, *socially responsive robots* and *sociable robots.* So long as the external approach fails to draw upon deeper models of human social competence, and, in particular, upon more realistic characterizations of impulses and emotions—internal social objectives, as Breazeal calls them—it will be impossible, she feels, to make robots that are genuinely social partners.

Other like-minded researchers observe that, although the artificial agents of the external robotics of emotion do exhibit true social competence, it is shown "only in reaction to human behavior, relying on humans to attribute mental states and emotions to the robot."[33] This amounts to saying that these agents do not have mental states, that they do not have true emotions, and that they are unable to feel

empathy. What passes for social relations with human agents depends on an ability to pretend, to simulate behavior—generally in a benign and fruitful way, but one that nonetheless rests on a mistaken perception on the part of their human partners. Indeed, much work on emotional behavior in robots, some of it significant, "has been driven by the fact that simple behavior may 'fool us' by leading us to perceive robots as having 'emotions' when, in reality, those 'emotions' are not explicitly modeled and just arise as an emergent effect of specific simple patterns of behavior."[34]

Robots can act *as if* they were afraid, for example, or *as if* they were aggressive, but in reality they do not have inner states equivalent to fear or anger, and there is nothing in them, no mechanism or module, that serves to produce emotions. As outside observers, we attribute emotions to robots to explain their behavior; but robots themselves do not have the emotions we attribute to them, for the excellent reason that they do not have inner states corresponding to them, nor do they have any machinery capable of producing them. Robots may well act *as if* they have emotions or *as if* they feel empathy, but they lack the internal dimension required to generate and justify these outward displays of sentiment. One therefore cannot treat the "emotions" expressed by these robots as real phenomena, only as illusory effects that occur as part of an interaction with a human partner, as phenomena that exist only in the eyes of an observer.

In effect, then, the idea that the emotions displayed by robots are not true emotions—that they amount to nothing more than pretending—functions as a methodological principle of the external approach. What is more, it has attracted interest outside cognitive science: not only actors and other performers, but painters and sculptors as well, actively assist and take part in research. This is unsurprising, really, for artists and roboticists share the same purpose: to provoke true emotions in reaction to emotions that, being the products of deliberate pretense, are in some sense false.

GALLERY 1

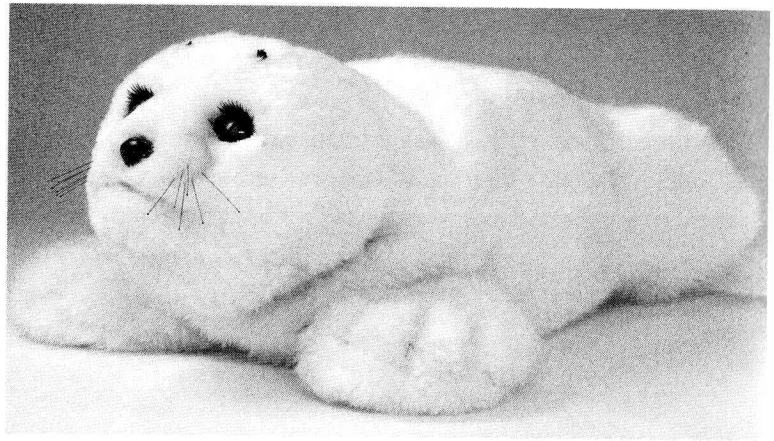
External Social Robotics

Keepon

Appearance: Doll-like
Expressive Modality: Noise; body movements:

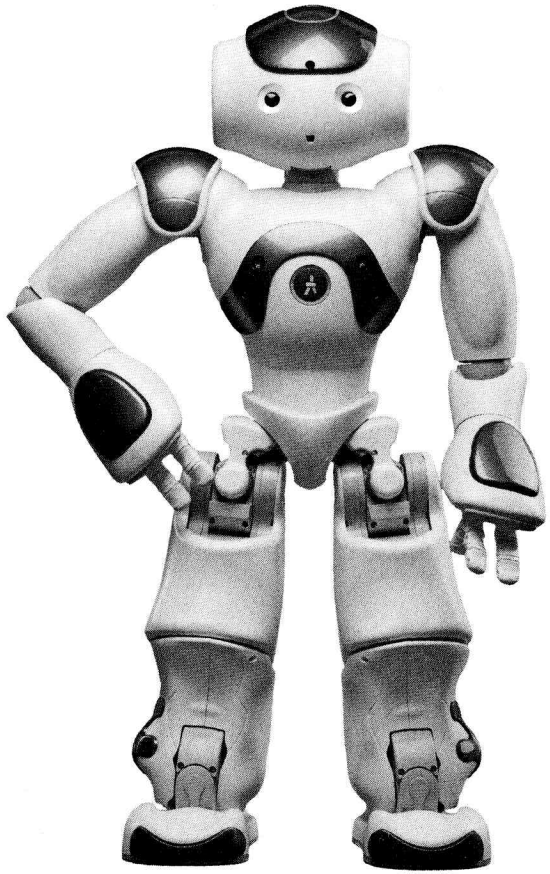
- lateral movements express pleasure
- vertical movements express excitation
- vibrations express fear

Receptive Modality: Tactile; visual (video camera)
Principal Use: Therapeutic mediator for autistic children; entertainment



Paro

Appearance: Animal-like
Expressive Modality: Cries; movements of body and eyelids
Receptive Modality: Tactile; word recognition; detection of loud noises and of the direction of the source
Principal Use: Therapeutic mediator associated with a variety of conditions: autism, dementia, depression; entertainment

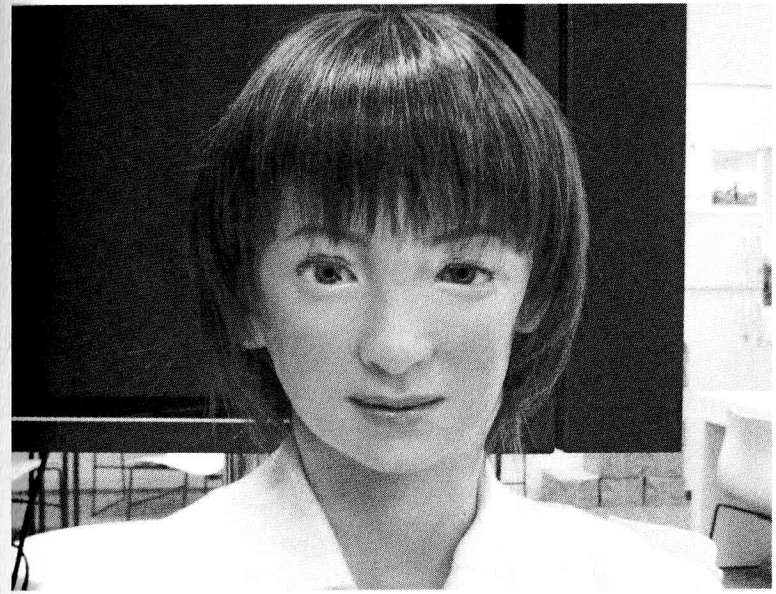


NAO

Appearance: Cartoon-like
Expressive Modality: Movements; posture; gestures; voice; audiovisual and proxemic signals
Receptive Modality: Tactile
Principal Use: Therapeutic mediator for autistic children and treatment of other developmental disabilities; education (teacher, instructor, coach); entertainment

KASPAR

Appearance: Child-like
Expressive Modality: Movements of the head, arms, hands, eyelids; posture; simple gestures; facial expression; voice
Receptive Modality: Tactile
Principal Use: Therapeutic mediator for autistic children

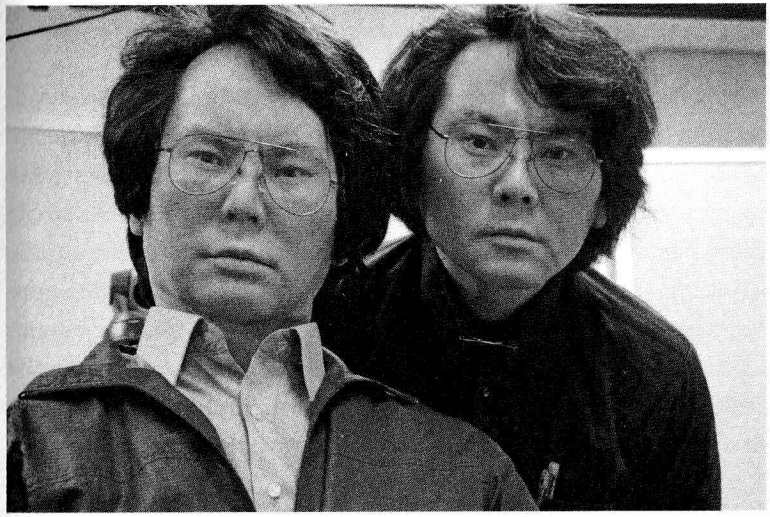


Saya

Appearance: Human-like
Expressive Modality: Realistic facial expression; posture; voice
Receptive Modality: Visual; aural
Principal Use: Education (teacher); receptionist

Face

Appearance: Human-like
Expressive Modality: Realistic affective facial expression (limited to joy, sadness, surprise, anger, disgust, fear)
Receptive Modality: Facial expression; eye movements
Principal Use: Therapeutic mediator for autistic children



Geminoids

Appearance: Human-like
Expressive Modality: Head movements; posture; voice
Receptive Modality: Visual; aural
Principal Use: Research on human-robot interaction; entertainment (theater)